Positive Selection Linked with Generation of Novel Mammalian Dentition Patterns

João Paulo Machado[1,2], Siby Philip[1,3], Emanuel Maldonado[1], Stephen J. O'Brien[4,5], Warren E. Johnson[6], and Agostinho Antunes[1,2,3,*]

[1]CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Porto, Portugal

[2]Abel Salazar Biomedical Sciences Institute (ICBAS), University of Porto, Porto, Portugal

[3]Department of Biology, Faculty of Sciences, University of Porto, Porto, Portugal

[4]Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russia

[5]Oceanographic Center, Nova Southeastern University, Ft Lauderdale

[6]Smithsonian Conservation Biology Institute, National Zoological Park, Front Royal, Virginia, USA

*Corresponding author: E-mail: aantunes@ciimar.up.pt.

Accepted: August 11, 2016

Data deposition: This project has been deposited at figshare under the inaccesso 10.6084/m9.figshare.3708099.

Abstract

A diverse group of genes are involved in the tooth development of mammals. Several studies, focused mainly on mice and rats, have provided a detailed depiction of the processes coordinating tooth formation and shape. Here we surveyed 236 tooth-associated genes in 39 mammalian genomes and tested for signatures of selection to assess patterns of molecular adaptation in genes regulating mammalian dentition. Of the 236 genes, 31 (~13.1%) showed strong signatures of positive selection that may be responsible for the phenotypic diversity observed in mammalian dentition. Mammalian-specific tooth-associated genes had accelerated mutation rates compared with older genes found across all vertebrates. More recently evolved genes had fewer interactions (either genetic or physical), were associated with fewer Gene Ontology terms and had faster evolutionary rates compared with older genes. The introns of these positively selected genes also exhibited accelerated evolutionary rates, which may reflect additional adaptive pressure in the intronic regions that are associated with regulatory processes that influence tooth-gene networks. The positively selected genes were mainly involved in processes like mineralization and structural organization of tooth specific tissues such as enamel and dentin. Of the 236 analyzed genes, 12 mammalian-specific genes (younger genes) provided insights on diversification of mammalian teeth as they have higher evolutionary rates and exhibit different expression profiles compared with older genes. Our results suggest that the evolution and development of mammalian dentition occurred in part through positive selection acting on genes that previously had other functions.

Key words: mammalian dentition genes, adaptive evolution, positive selection, tooth-associated genes, teeth.

Introduction

As a major determinant of vertebrates' ecology, teeth have played a crucial role in species survival. Teeth have been subjected to strong selective constraints because they first appeared in the oral cavity in jawed vertebrates over 460 Myr during the Ordovician (Smith and Coates 1998). While mammalian teeth share basic components, they exhibit great diversity in number, size and shape (fig. 1*A*). However, in spite of their importance for animal survival, teeth have been lost independently in multiple lineages of tetrapods' (Davit-Beal et al. 2009), including mammals (e.g., pangolins). And others mammals have teeth with little or no enamel (e.g., sloths) or have booth teeth and enamel reduction (e.g., platypus).

Mammals differ from other living vertebrates by having very complex teeth and a restricted capacity for tooth renewal (Jernvall and Thesleff 2012). Moreover, in mammals there is a strong correlation between feeding habits, patterns of tooth formation (e.g., cardiform, villiform, incisor, canine, molariform) and their number of teeth (Koussoulakou et al. 2009)

(fig. 1*A*). While some nonmammals have multi-rowed dentition and replace their teeth regularly throughout their lifetime, mammals have only one row of teeth and either renew their teeth only once or in some rodents without any replacement (Jarvinen et al. 2009; Koussoulakou et al. 2009; Mikkola 2009). Thus, vertebrate evolution is characterized by a reduction in tooth number (from polyodonty to oligodonty), by a shift in timing of tooth development (from polyphyodonty to di- and/or monophyodonty) and by an increase in morphological complexity (from homodonty to heterodonty) (Salazar-Ciudad and Jernvall 2004). In addition, these mammalian features, including increased shape complexity, multi-cusp teeth, and stable tooth number, facilitated the maintenance of the high metabolic rates of mammals by ensuring efficient processing of food (Armfield et al. 2013).

Modern mammalian dentition develops through a series of well-defined morphological stages that require sequential and reciprocal interactions between the epithelium and mesenchyme tissues (Mitsiadis and Graf 2009). In mice, the first sign of tooth development, the thickening of the oral epithelium, is observed at embryonic day 10.5 (E10.5) (Zhang et al. 2005; Mitsiadis and Graf 2009), when tooth sites and types are established (Zhang et al. 2005). Between embryonic days 12.5–13.5 (E12.5–E13.5) the tooth bud is progressively formed following the epithelium invagination of the underlying mesenchyme (Mina and Kollar 1987; Mitsiadis and Graf 2009). During days 14.5–15.5 (E14.5–E15.5) the growth of the epithelium leads to the formation of the cap structure (Mitsiadis and Graf 2009) and to its configuration during days 16.5–18.5 (E16.5–E18.5) (Mitsiadis and Graf 2009). During the late bell stage, embryonic day 18.5 (E18.5), mesenchyme cells form the dental follicle and dental pulp (Mitsiadis and Graf 2009) (fig. 1*B*).

In spite of the wide phenotypic diversity among mammal dentition patterns, previous studies have demonstrated only slight differences in gene expression patterns, with human and mice teeth sharing considerable homology in ontogenesis and underlying molecular networks (Lin et al. 2007). The marked similarity between odontogenesis (in lamina, bud, cap, and bell stages) and gene expression profiles (Zhang et al. 2005) in mice and humans suggests that there are strong functional constraints in mammalian teeth development. Genetic control of tooth development encompasses, to-date, more than 300 genes (Thesleff 2006). However, this is probably an underestimate, because analyses of large data sets and new approaches using microarray profile search functions have identified additional genes associated with odontogenesis (Kim et al. 2012; Landin et al. 2012).

Genes involved in adaptation and functional innovation often show the footprints of positive selection through elevated ratios of nonsynonymous to synonymous nucleotide substitutions (Yang and Bielawski 2000; Nielsen et al. 2005;

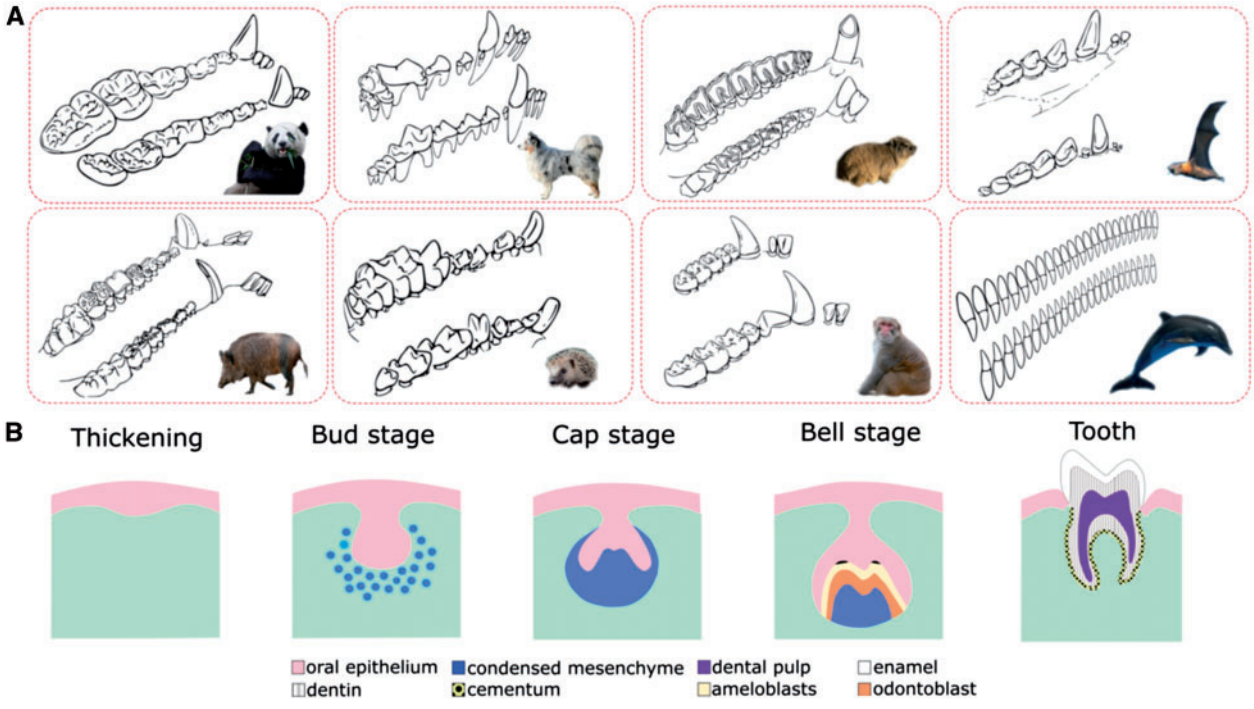

Fig. 1.—Phenotypic diversity and developmental stages of mammalian dentition. (*A*) Examples of the phenotypic diversity in mammalian dentition, presented clockwise the images of upper and lower dentition in giant panda, dog, pika, megabat, dolphin, macaca, hedgehog and pig (images adapted from Hillson 2005). (*B*) Typical mammalian tooth developmental stages (image adapted from Volponi et al. 2010).

Philip et al. 2012). Here we performed comparative evolutionary analyses of tooth-related genes to identify signatures of selection that may have shaped tooth phenotypic diversity among mammals. Of the 236 tooth-associated genes analyzed in 39 mammalian genomes, we detected strong selection signatures in 31 genes using both gene- and species-trees. Moreover, younger genes (mammalian-specific) had accelerated evolutionary rates, and differential expression profiles when compared with older genes (vertebrate-specific).

Materials and Methods

Sequences and Annotation

Genes associated with tooth development, tooth disease and mammalian tooth phenotypes were retrieved from the Gene Ontology (GO) database (Ashburner et al. 2000; Harris et al. 2004) and the Rat Genome Database (RGD) (Shimoyama et al. 2011; Laulederkind et al. 2013). To restrict the gene data set, we only used the associated processes listed in the GO and Mammalian Phenotype (MP) (supplementary table S1, Supplementary Material online). The final data set included 247 genes, from which 11 genes were later excluded as fewer than 20 sequences were available. For the 247 genes we obtained 7,892 coding sequences corresponding to a unique transcript, when available, for each species. The sequences used in this work were retrieved from *ENSEMBL* v64 or v65 (Flicek et al. 2012) using PyCOGENT 1.5.3 (Knight et al. 2007) implemented in EASER (Maldonado et al. 2013) querying *ENSEMBL COMPARA* database. All the retrieved results were manually inspected and when the sequences could not be retrieved using the script, they were manually downloaded. The corresponding gene coordinates were obtained using *BIOMART* in *ENSEMBL* to construct the annotation file needed to build the ideogram in Idiographica (Kin and Ono 2007).

Gene Tree-Based Reconstruction

For each gene a multiple sequence alignment (MSA) was built using the retrieved coding sequences translated to amino acids and further back-translated to nucleotides and MUSCLE (Edgar 2004) implemented in SEAVIEW v4 (Gouy et al. 2010). The MSAs were refined in GBLOCKS (Castresana 2000) using the relaxed parameters (Talavera and Castresana 2007) to reduce the false positives resulting from improper aligned positions. The filtered MSA was used to inspect possible evolutionary models using MrAIC (Nylander 2004). We restricted to Bayes models to save calculation time and used AICc (Akaike information criterion correction) for models comparison. Phylogenetic gene-based tree reconstructions were obtained with PhyML v3.0 (Guindon et al. 2009) under the previously estimated evolutionary model and the topology branches support values were retrieved using the aLRT (Approximate likelihood-ratio) test (Anisimova and Gascuel 2006). The tree topology was further used as the gene tree in evolutionary analyses after the removal of branches length, allowing CODEML to calculate each branch length during the likelihood estimation of each model. The final data set incorporated 236 filtered alignments (corresponding to 236 genes), obtaining an average of 33.44 sequences and length ~704.12 bp per MSA. The species tree topology was obtained from *ENSEMBL* (supplementary fig. S1, Supplementary Material online). Trees were pruned, as necessary due to missing taxa, using Phyutility (Smith and Dunn 2008).

Evolutionary Rate and Protein Age

For each gene the number of nonsynonymous substitutions per nonsynonymous site (d_N) and the number of synonymous substitutions per synonymous site (d_S) were calculated using a maximum-likelihood method CODEML implemented in PAML v4.6 (Yang 2007). Estimations of d_N, d_S and d_N/d_S, were obtained using six different models (Model 0, 1a, 2a, 7, 8 and 8a). Equilibrium codon frequencies of the model were used as free parameters (CodonFreq = 2). Model 0 (M0, one-ratio) was used to estimate global d_N/d_S, d_N and d_S. Model 1a (M1a, nearly neutral) distributes the sites in two site-classes varying between 0 and 1, assuming that all sites have $d_N/d_S \leq 1$. Model 2a (M2a, positive selection), unlike M1a, estimates the proportion of sites under positive selection, $d_N/d_S > 1$. Models 7 (M7, beta) and 8 (M8, beta + $\omega > 1$), approximate the d_N/d_S variation over sites through a beta distribution, estimating the proportion and the d_N/d_S ratio of the positively selected sites, whereas M8 only includes site-classes above neutrality. The models allowing positive selection along the alignment (M2a and M8) were compared pairwise against stricter models, M1a and M7, respectively, using likelihood ratio tests (LRT). Each calculation of the LRT corresponds to $2\times[lnL$ (alternate model)$-lnL$ (*null* model)] (or $LRT = 2\times(\Delta lnL)$). Comparisons between models M8 and M8a were used to identify deviations from neutrality. This pairwise comparison focuses on testing whether sites belonging to a site-class with a $d_N/d_S > 1$ are evolving differently from near neutrality ($d_N/d_S \approx 1$). For each pairwise comparison, M1a versus M2a, M7 versus M8, M8 versus M8a, the LRT obtained were compared against a χ^2 distribution. The degrees of freedom, used to obtain the χ^2 critical values, were the difference in the number of parameters in the *null* and *alternate* model for each pairwise test. The results from CODEML were corrected for possible multiple testing bias using the procedure of Benjamini and Hochberg (Benjamini and Hochberg 1995) as implemented in the program Q-Value (Storey and Tibshirani 2003). For each *P* value, we also estimated the corresponding *q* value. When the *q* value was below, the *P* value obtained for the LRT value the gene was considered to be under positive selection (1), and when above, the gene was considered negatively selected (0).

The positions of the positively selected sites were mapped to the human sequences using an in house script (available upon request). Because positive-selection analyses tend be less reliable in regions of poor alignment, for quality control all MSA used for testing for positive selection were submitted to GUIDANCE (Penn et al. 2010) to obtain an alignment confidence score. The correlation and the confidence estimates of each alignment were plotted in a scatter plot, which showed no link between the confidence estimates of the alignment and the positively selected sites (supplementary fig. S2, Supplementary Material online).

Exons and Introns Substitution Rate

Gene coordinates obtained from ENSEMBL BIOMART were used to retrieve the phyloP (Pollard et al. 2010) site scores for introns and exons using the USCS browser (Kent et al. 2002). The pre-calculated values available in USCS table phyloP44wayPlacMammal that were used only included placental mammals (Goldman et al. 2013) and the values were obtained using the coordinates of reference sequences from human (hg18). The empirical cumulative distribution function (ECDF) from introns and exons of phyloP scores and the Mann–Whitney *U* values were obtained using MATLAB vR2014b. Given that the number of analyzed positions (intronic and exonic) from negatively selected genes was greater than the number of positions in positively selected genes, we built a script for sampling (allowing repetitions) the values from each of the intronic and exonic regions of the negatively selected (conserved) genes. For comparisons between positively and negatively selected genes, and to diminish calculation times, both pools of values were restricted to: (1) 1,000,000 points in introns, (2) 100,000 sampling points in first introns, and (3) 100,000 points in exons. To validate the procedure, for each scenario three random samples of introns and exons from positively and negatively selected genes were generated from each pool of values and were tested for homogeneity using the Mann–Whitney *U* values.

Protein Age, Characteristics and Functional Clustering

Protein ages were estimated with PPODv4_Ortho MCL_families and *Dollo* parsimony and grouped into three age classes defined as: $\leq$220.20 Myr (Mammalian specific), >220.20 Myr and $\leq$454.60 Myr (Vertebrate specific), >454.60 Myr (Older proteins) using *ProteinHistorian* (Capra et al. 2012) (supplementary table S2, Supplementary Material online). For positively selected genes, the disorder status was calculated for each protein with SPINED (Zhang et al. 2012) using human sequences as the query. Positively selected genes were grouped into functional clusters based on DAVID (Huang da et al. 2009). The protein interactions were retrieved from BioGRID (Stark et al. 2006; Chatr-Aryamontri et al. 2013) and all proteins with more than 100 interactions were excluded. Statistical analyses were performed in SPSS v20.

Expression of Tooth-Associated Genes during Development

Expression profiles were obtained from NCBI GEO (Barrett et al. 2005, 2013). We used two experiments from mouse corresponding to: (1) tooth germ tissue at embryonic day 13.5 (Lachke et al. 2012) [GEO:GDS4453] and, (2) postnatal stage (Pemberton et al. 2007) [GEO:GSE7164] and one experiment corresponding to embryonic stages from 4 to 9 weeks after fertilization in humans (Yi et al. 2010) [GEO:GSE15744]. For each data set the GPL-associated (GEO Platform) files were used to filter the tooth-associated genes, their expression values were log_2 normalized to reduce numerical noise. For the different probes associated with the same gene, their values were averaged. Cluster analysis using k-means was performed in MATLAB vR2014b and the statistical analysis was performed using SPSS v20.

Results

Gene Localization and Functions

Genes associated with tooth development were plotted on an ideogram to show their location in the human genome (fig. 2). Of the 247 tooth-associated genes, 10 are located on Chromosome (Chr) X and one on Chr Y, whereas the remaining 236 are autosomal. *MECP2* was the only X-linked gene with evidence of positive selection, compared with 30 positively selected autosomal genes. For molecular evolution analysis, such as CODEML, the 247 were reduced to 236 genes, because we were unable to retrieve more than 20 orthologs for these 11 genes.

The majority of the tooth-associated genes identified in this study are also involved in other processes and therefore they were not restricted to tooth-associated processes (supplementary table S1, Supplementary Material online). For example, a pleiotropic effect has been reported for genes such as *BMP4*, which is primarily associated with colorectal cancer (Houlston et al. 2008) but also Parkinson's disease (Simon-Sanchez et al. 2009).

Selective Regimes in Tooth-Associated Genes

When the gene trees were used as input for CODEML analysis, M8 was significantly more adjusted in 148 genes relatively to M7 (supplementary table S3, Supplementary Material online), although when using the strict pairwise comparison M8 versus M8a, only 35 genes showed that the site class was ($\omega > 1$) significantly above neutrality. When using the species tree, selection analyses supported the alternate model (M8) in 160 genes, whereas in the M8 versus M8a comparison, 48 genes favored M8. 31 genes (~13.1%) showed signatures of positive selection in both analyses (gene tree and species tree)

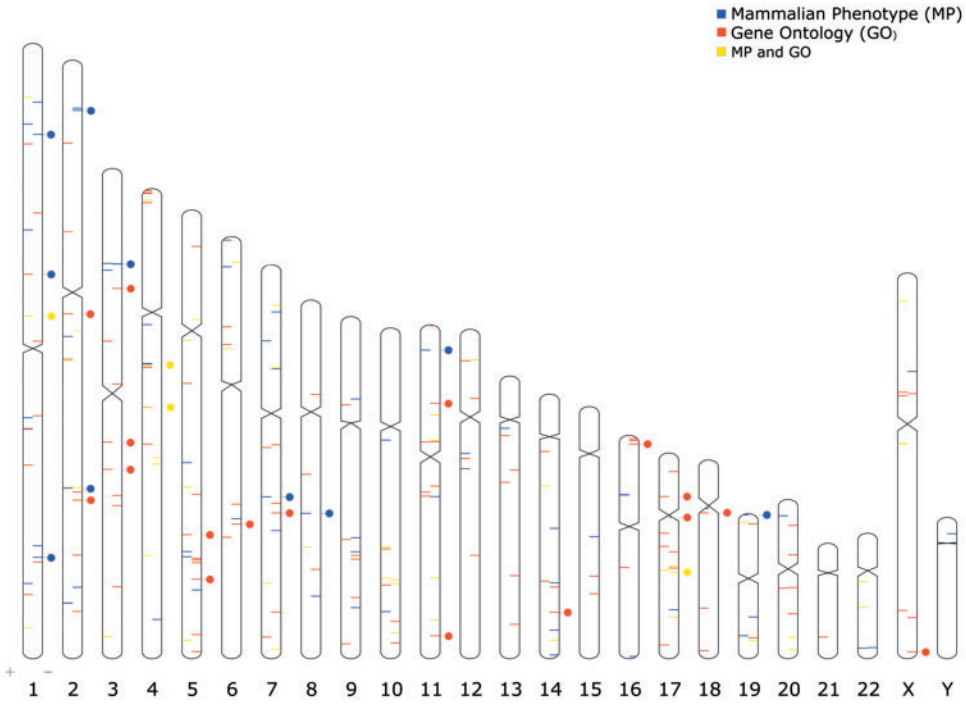

Fig. 2.—Ideogram of the human genome. Human chromosomal location of tooth-associated genes. Each chromosome is labeled with its respective number (autosomal chromosomes) or letter (sexual chromosomes) code. The symbols (+) and (−) represent the DNA strand orientation. The circles near the gene represent significant evidence of positive selection.

(supplementary table S4, Supplementary Material online). The comparisons between gene tree and species tree allowed us to reduce the possible bias that an incorrect phylogenetic topology may have introduced into positive selection analyses.

Pairwise comparisons of M7 versus M8 have previously been shown to be less robust (but more powerful) than M1a versus M2a comparisons (Nielsen and Yang 1998). Under model M2a and using the gene-based tree, 28 genes had signatures of positive selection (supplementary table S5, Supplementary Material online), whereas the alternate model was favored in 37 genes when the species tree was used (supplementary table S6, Supplementary Material online). Although the analyses using M2a versus M1a is significantly faster to run the analysis, this comparison retrieved 20 of the same genes identified as being under positive selection using pairwise M8 versus M7 and M8 versus M8a comparisons. In aggregate, 20 genes were identified with signatures of selection regardless of the model and phylogenetic assumption used. Despite being a more-rapid model, M2a was the most sensitive model to the phylogenetic assumptions because the results obtained from the species tree and gene tree were less similar when compared with the more parameter-rich pairwise analysis (supplementary fig. S3, Supplementary Material online). The Spearman's correlation between the model M2a versus M1a and M8 versus M7 showed that the primer model comparison is more sensitive to the choice of input tree used in the detection of positive selection (supplementary fig. S3, Supplementary Material online).

Using M8 with the 31 positively selected genes and the gene tree as parameters, 236 positively selected sites were identified. These were compared with 235 sites identified using the species tree. Using the same approach (i.e., concordance between species and gene tree), we identified 181 sites under positive selection (posterior probability >0.95) independent of the phylogenetic assumption. The positions of the positively selected sites were annotated using the human protein as reference (supplementary table S7, Supplementary Material online). The posterior probabilities were calculated for each site using human sequences as references for M8 results (using gene-tree as phylogenetic tree). These data show that these positively selected sites are distributed randomly and are not concentrated on the ends of the genes (supplementary fig. S4, Supplementary Material online). Remarkably, ~67% of the positively selected sites were located in disordered regions, based on the results from SPINE-D (Zhang et al. 2012), and therefore correspond to regions that commonly have a less stable tertiary structure (fig. 3).

Alignment Uncertainty and Phylogenetic Resolution

The MSAs from the positively selected genes were submitted to GUIDANCE to confirm that alignments were robust and therefore that most of the positive selection was not due to improper alignment or to uncertainty in some regions. In the

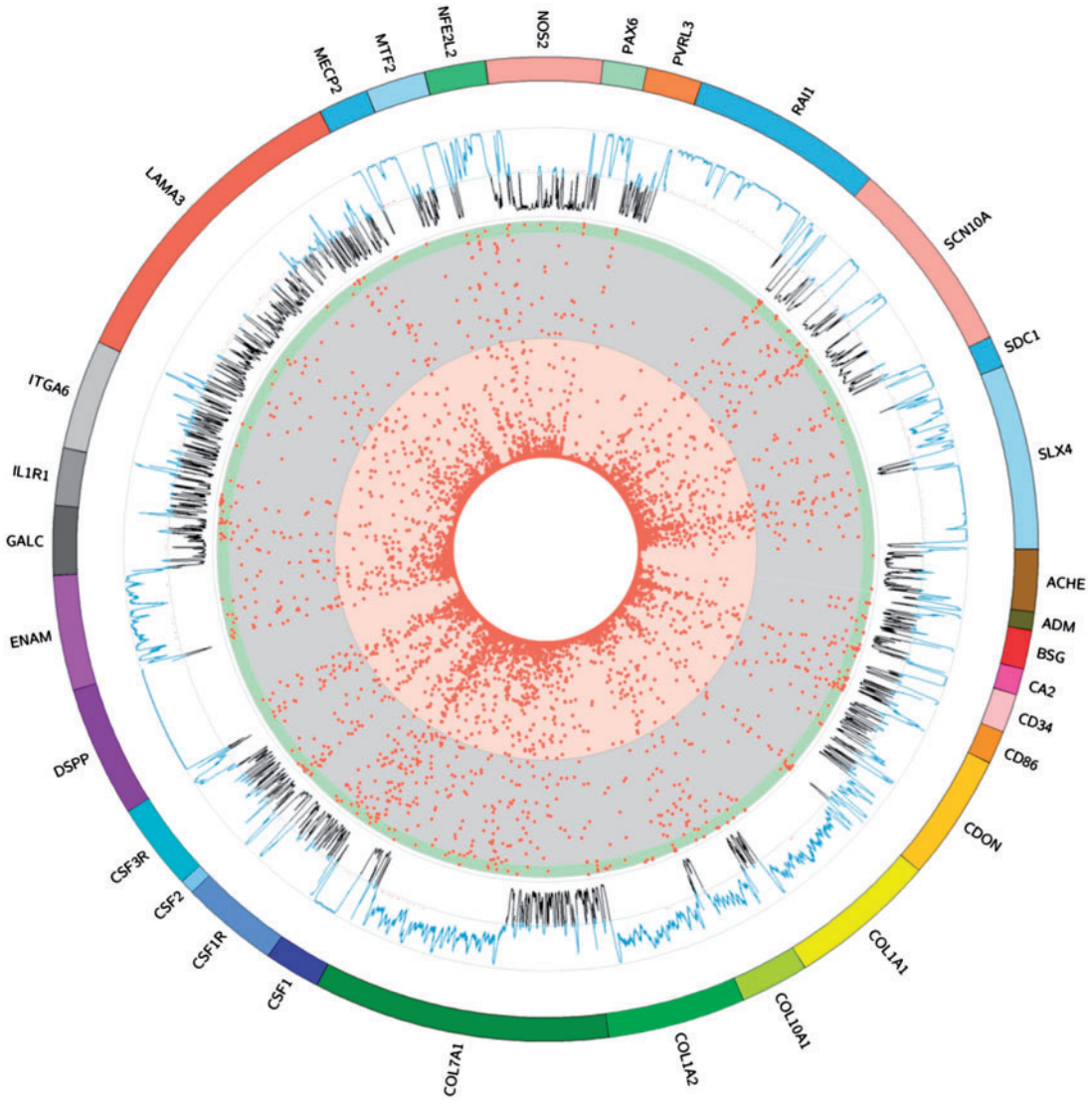

Fig. 3.—Tooth-associated genes under positive selection. The Bayesian Empirical Bayes posterior probability under M8 obtained using the gene tree is plotted in red dots in the center of the figure. The dots in the inner circle correspond to a PP from 0 (interior) to 1 (outer). The graphic line corresponds to the calculated disorder probability. The blue lines corresponding to values >0.5 are considered intrinsically disordered regions.

31 positively selected genes, no associations were observed between the proportion of sites under selection and any detected uncertainty in the alignment (supplementary fig. S5, Supplementary Material online). Because the terminal portions of alignments tend to be more difficult to align, it has been reported (Markova-Raina and Petrov 2011) that these regions may have high false-positive ratios. However, in our data set the positive-selected sites were dispersed relatively evenly from tail to core, decreasing the probability that poor alignment quality led to some false-positive or false-negative results. Moreover, the TREE-PUZZLE results showed that there was no association between evolutionary rate and uncertainty in the phylogenetic signal, as the majority of the positively selected genes had <10% of unresolved quartets (with the exceptions of *ADM*, *AQP6*, *CA2*, *CSF2*, *MTF2*, and *PVRL3*) (supplementary table S8, Supplementary Material online).

Intronic Acceleration in Positively Selected Genes

The ECDF analyses showed that positively selected genes had accelerated rates in both exonic and intronic regions (fig. 4*A* and *B*) when compared with the negatively selected genes scores (fig. 4*C* and *D*). In the introns or exons of positively selected genes, there was a significantly higher departure from neutrality when compared with negatively selected genes (Mann–Whitney *U* test, $P < 0.01$). This result is consistent with the observation that over 50% of the more accelerated sites were within lower phyloP scores (fig. 4A–*D*).

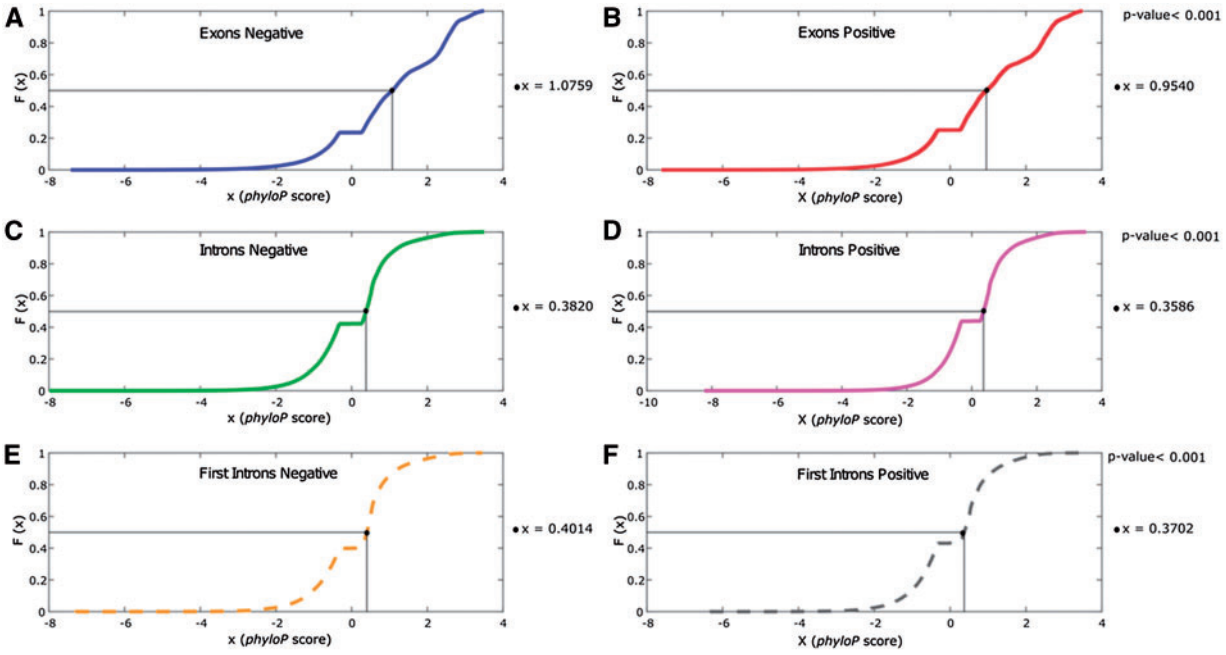

Fig. 4.—Comparison between *phyloP* scores of positively and negatively selected genes. Empirical Cumulative Distribution Function obtained for tooth-associated genes, introns and exons. The *x* corresponds to phyloP scores values and F(*x*) corresponds to the cumulative frequency. *P* values correspond to the significance of Mann–Whitney *U* test result from the three pairwise comparisons. (*A*) Exons from non-positively selected genes, (*B*) Exons from positvely selected genes, (*C*) Introns from non-positively selected genes, (*D*) Introns from positvely selected genes, (*E*) First introns from non-positively selected genes and (*F*) First introns from positively selected genes.

Although these phyloP scores were obtained from USCS computed values excluding the nonplacental mammals, there is no expectation that this would significantly change the interpretations of the phyloP analysis. The first intron of positively and negatively selected genes also was significantly different (*P* value < 0.001) (fig. 4*E* and *F*), which is evidence of a consistent higher evolutionary rate of introns concordant with the presence of positive selection in coding regions.

The proteins were classified into three distinct phylogenetic groups according to their predicted gene age as Mammalian (mammalian-specific proteins), Vertebrate (vertebrate-specific) and Old (old proteins). An analysis of the binning patterns of the proteins based on their estimated ages revealed that introns of vertebrate-specific genes were the most accelerated while exons of genes classified as "old" were more conserved (table 1). The phyloP analyses of introns and exons (gene structure) of the genes of different ages suggested that the most accelerated groups ranked (highest to lowest): vertebrate introns > vertebrates 1st intron > older introns > older 1st intron > mammal's introns > mammals 1st intron > mammal's exons > vertebrate's exons > older exons. This pattern revealed that mammalian-specific genes had the mostaccelerated exons and the least accelerated (most conserved) introns compared with vertebrate-specific and "older" genes.

Positively Selected Genes Implicated in Diseases

Of the 31 genes under positive selection, 21 have a hypothesized or known phenotypes associations in OMIM database. However, only four of these, *COL1A1*, *COL1A2*, *DSPP* and *ENAM*, have phenotypes that have been specifically associated with teeth. Proteins col1a1 and col1a2 are associated with osteogenesis imperfecta type I, dspp is associated with dentin dysplasia type II, dentinogenesis imperfecta shields type II and dentinogenesis imperfecta shields type III, and enam is associated with amelogenesis imperfecta type IB. The functional clustering analysis, using a classification stringency of "*high*", identified 16 clusters from the 31 positively selected genes (supplementary table S9, Supplementary Material online). Two of these clusters were associated with biomineralization and/or structural constituents of tooth enamel (*ACHE*, *COL1A1*, *DSPP* and *ENAM*).

Acceleration of Recent Proteins

For each age-dependent protein cluster we estimated average omega, number of positively selected sites and GC content and searched for associated GO process. Despite high variability, d_N/d_S estimates from M0 in CODEML supported the hypothesis that more-recently evolved proteins had accelerated evolutionary rates (fig. 5*A*), as the average omega from mammalian-specific proteins was slightly higher than proteins that

Table 1

Pairwise Comparison between Introns and Exons Categorized by Age

		Mammals			Vertebrates			Older		
		Exons	1st Intron	Introns	Exons	1st Intron	Introns	Exons	1st Intron	Introns
Mammals	Exons	–	1	1	0	1	1	0	1	1
	1st Intron		–	1	0	1	1	0	1	1
	Introns			–	0	1	1	0	1	1
Vertebrates	Exons				–	1	1	0	1	1
	1st Intron					–	1	0	0	0
	Introns						–	0	0	0
Older	Exons							–	1	1
	1st Intron								–	0
	Introns									–

NOTE.—The values were obtained in ranksum test in MATLAB, when the value is 1, the left entry is more show lesser acceleration, whereas when the value is 0 suggest acceleration of the left entry relatively to the top entry (e.g., exons aged as mammalian specific are more accelerated than exons in vertebrate-specific, but lesser accelerated than 1st introns in mammalian specific genes). The three different structures exons, first intron and introns are compared in three age classes, mammalian-, vertebrates-specific and all the other predicted ages are categorized as older proteins.

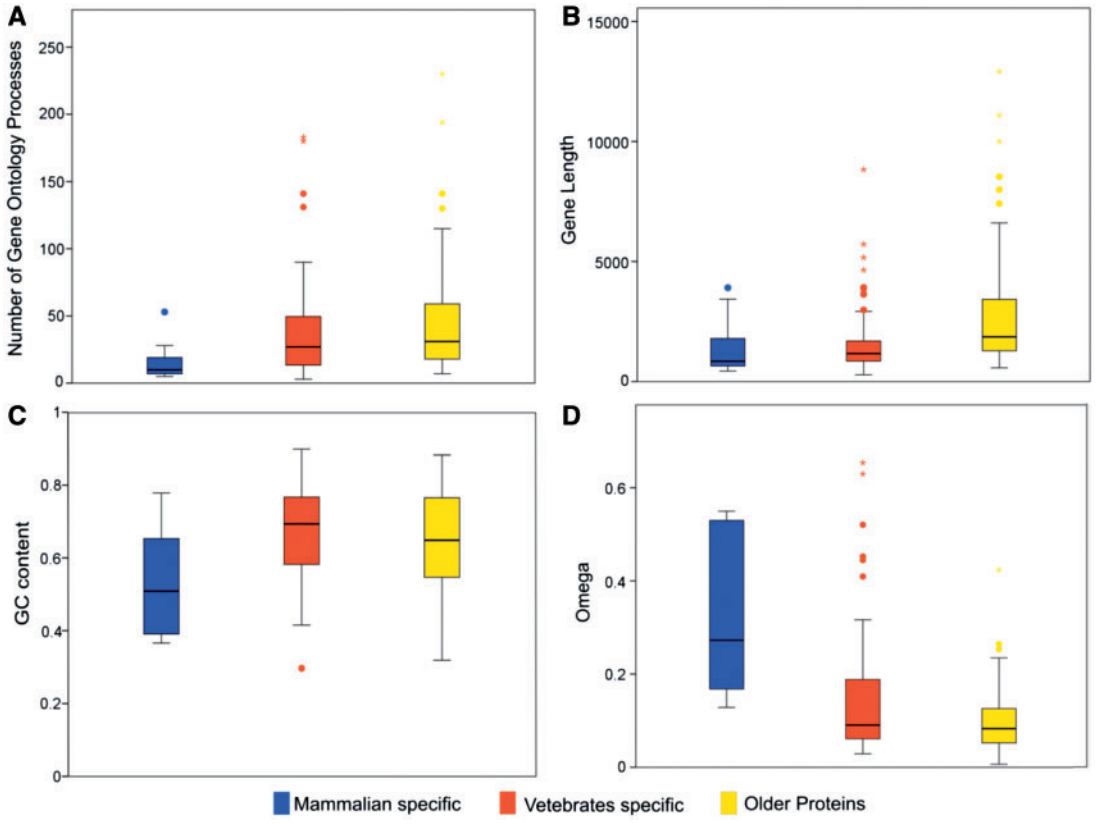

FIG. 5.—Protein age and tooth-associated genes. Relation between estimated protein age, classified as mammalian, vertebrate and older proteins and (A) the GO number (Gene Ontology) processes, (B) gene length, (C) GC content and (D) evolutionary rate (Omega).

arose before the mammalian divergence. The younger proteins, that is, mammalian specific, were shorter, were related with fewer GO terms, had protein coding sequences with slightly lower GC content, and had fewer interactions (fig. 5*B–D*).

Expression Pattern of Tooth-Associated Genes

Expression data supported the hypothesis that the younger genes are less expressed in early stages of tooth development. The GDS4453 experiment, which corresponds to a primary stage of tooth development in mice E13.5, showed that at

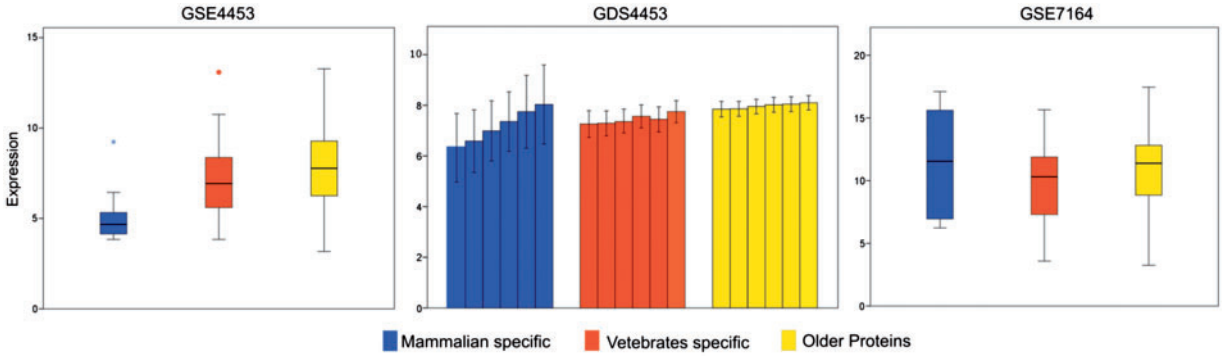

Fig. 6.—Expression profile of tooth-associated genes. Results from experiments GDS4453, GDS4453 and GSE7164 are represented from left to right, respectively. In GDS4453, the expression levels are grouped according to the estimated age and from 4th to 9th week (left to right). The mammalian-specific tooth-associated protein-coding genes are down-regulated in early developmental stage.

this stage there is a slightly lower expression of young proteins. Moreover, results from GSE7164 (fig. 6), which corresponds to a postnatal stage, showed that there is a more similar expression pattern of the younger proteins compared with either vertebrate-specific or "old" proteins.

The expression data from GDS4453, corresponding to weeks 4–9 of human embryonic development, revealed that the expression of younger proteins was lowest from the 4th to 6th week, similar to the patterns observed in other stages (GDS4453 and GSE7164). Interestingly, the 31 positively selected genes had different expression patterns during these stages, because of the 16 k-clusters examined, only clusters 1, 5 and 6 did not have any genes with patterns of positive selection (supplementary fig. S6, Supplementary Material online).

Discussion

Most vertebrates possess teeth in jaws, with a few exceptions, including birds, which lost their teeth through evolution. Therefore, as teeth first appeared in jawed vertebrates ~460 Ma (Smith and Coates 1998), dentition has been subjected to purifying selection. The appearance of teeth involved an intricate coordination of multiple genes that likely shared in the functions that were required for the coordination of tooth development. However, most of these were not novel genes, but had previously had other functions. Genes that are physically located close to each other chromosomally are more likely to be co-expressed and to share a common ancestral function than more dispersed genes (Cohen et al. 2000; Woo et al. 2010). However, tooth-associated genes in the mammalian genome are widely dispersed (fig. 2). This suggests that tooth development depends on the coordination of multiple genes that previously were involved in a variety of different functions. The earliest tooth-like structures of the vertebrate oral cavity were first located outside the mouth and served diverse functions including protection, sensation and hydrodynamic advantage (Koussoulakou et al. 2009).

While the majority of the genes studied have been subjected to purifying selection, we identified 31 genes that evolved under positive selection, with specific sites with d_N/d_S significantly >1. Previous studies have shown that positively selected sites are functionally relevant (Morgan et al. 2012; Dasmeh et al. 2013) and therefore sites with d_N/d_S significantly above one are expected to have a determinant fitness role. Since natural selection has shaped the current diversity of tooth dentition in mammals, sites with evidence of positive selection signatures should be linked with differential selective advantages in each species. However, distinguishing neutral selection from a positive-selection regime acting on genes is often complicated. Here we overcome this uncertainty by comparing the results from two robust analytical approaches (using gene trees and species trees) to detect selection, with the premise that this dual approach is more reliable and less subject to statistical noise when estimating the degree of selective pressures acting on the genes.

While the majority of the identified sites were evolving under negative selection, the presence of sites with a $\omega > 1$ supports their role in determining protein functionality, and therefore demonstrates their role in the development of mammalian phenotypic differentiation. These results demonstrate that tooth-associated genes have different selection signatures and therefore affirms their important role in mammalian adaptations. We identified 31 genes that are most-likely responsible for the tooth diversification among mammals. Within these 31 genes, we found 181 sites under positive selection, most of which were located within intrinsically disordered protein regions. This confirms previous findings that there is an over representation of positively selected sites encoding intrinsically disordered regions of proteins (Nilsson et al. 2011). Furthermore, there was no evidence of underrepresentation of functional amino acids in intrinsically disordered regions of proteins (Nilsson et al. 2011). Positive

selection in tooth-associated genes was more persistent in disordered regions, which is important since disordered regions allow proteins to access target sequences and influence local conformation and activity (Collins et al. 2008). Moreover, there is a strong correlation between biomineralization and structural disorder of proteins (Kalmar et al. 2012). Therefore, these sites, particularly those corresponding with disordered regions, are potentially of prime relevance to the function of these proteins, and thus are potential sites for further site-directed mutagenesis studies.

Within the group of positively selected genes we identified two clusters of genes that were involved in tooth-specific processes, biomineralization and structural organization of tooth specific tissues. Because these two gene clusters were composed of ones that have been identified as being crucial for tooth formation, they are potential candidates for future study to determine their specific roles in the phenotypic diversification of the dentition in mammals. For example, one of these positively selected genes, *ENAM*, was previously demonstrated to have signatures of positive selection in human populations (Kelley et al. 2006) and in Kalmar dogs (Kalmar et al. 2012). Although *ENAM* has been previously been characterized as a multifunctional protein that is essential in early stages of tooth development (Landin et al. 2012), our re-analysis of data from three different microarrays (Pemberton et al. 2007; Yi et al. 2010; Lachke et al. 2012) suggest that there is a higher expression rate of mammalian-specific genes such as *ENAM* during tooth development in later stages. *ENAM* has also been linked with tooth enamel thickness and dietary changes in primates (Kelley and Swanson 2008). Our analyses also suggest that *ACHE*, *COL1A1* and *DSPP* have been involved in mammalian dentition adaptations.

Previous studies have demonstrated a high degree of sequence conservation in introns (Hare and Palumbi 2003) and among intron positions in orthologous genes (Henricson et al. 2010), and have observed that regions under negative selection, known as mutational cold spots, often correspond to regions that are more negatively selected than protein coding regions (Katzman et al. 2007). Concordantly, introns in negatively selected genes are also under a higher selective regime than in positively selected genes. Given the functional importance of the intronic regions, it is expected that this asymmetrical evolutionary rate may have functional relevance. It has also been demonstrated that changes in noncoding regions are associated with rapid evolutionary changes in enamel thickness and that they can have a major impact through differentially altering the affinity of transcription factors that regulate tooth development (Horvath et al. 2014). These mammalian intronic regions (especially the first intron) often have regulatory elements (Oshima et al. 1990; Jonsson et al. 1992). In slight contrast, here, we also observed a evolutionary patterns in both the first introns of the positively selected genes and of negatively selected genes. Our results provide further support that purifying and positive selection can have a strong effect on intron sequence evolution, as was observed between humans and chimpanzees (Gazave et al. 2007).

The dental gene network core has been a common feature of all species because the first species with pharyngeal teeth and including all of its jawed descendants (Fraser et al. 2009). This dental pattern has been associated with an ancient dental regulatory network (*BARX1*, *EVE1*, *LHX7*, *LHX8* and seven HOX's genes) and a dental circuit (*BMP2*, *BMP4*, *DLX2*, *EDA*, *EDAR*, *PAX9*, *PITX2*, *RUNX2* and *SHH*) that has also been reported in cichlids (Fraser et al. 2009). Our analyses have highlighted that this ancient suite of genes, except *EVE1* (which was not included in this study), have evolved under purifying selection. This confirms the hypothesis that this core dental network is "evolutionarily essential" because there is no corrected patterning of the dentition without the involvement of those genes and the appearance of those genes predates the vertebrates' emergence (Fraser et al. 2009).

Our results extend previous reports of correlations between evolutionary rate, structural properties and age class (Toll-Riera et al. 2012) and provide evidence that younger proteins (mammalian-specific proteins) are involved in fewer GO processes, are involved in fewer interactions, are shorter and have higher evolutionary rates. Similarly, GC content in these younger proteins is slightly lower than in older protein-coding sequences. Although some of these observations have been previously reported, the importance of these patterns is still being debated. In our data set, higher evolutionary rates were observed in the younger proteins, suggesting that most of the phenotypic diversity observed in the mammalian dentition may rely on "new proteins", whereas "older" proteins are more-likely to be under strong purifying selection. In addition, our analyses of expression data revealed that these younger proteins are expressed less in early stages of tooth development compared with later stages.

Conclusions

We conducted a top–down analysis of 236 tooth-associated genes and our results revealed 31 genes with evidence of significant positive selection. Positively selected sites tended to be located in disordered regions of the protein, and therefore are more likely to be functionally relevant. Clustering analysis identified four genes (*ACHE*, *COL1A1*, *DSPP* and *ENAM*) with signatures of positive selection and which are associated with odontogenesis. However, their role in the diversification of mammalian phenotypes is still unknown. The asymmetrical evolutionary rate among introns of the positively selected genes and the negatively selected genes suggests that intronic regions may also have had a role in mammalian diversification. Age-class analyses revealed that more-recently evolved proteins are expressed in later developmental stages

and, given their higher evolutionary rate, are probably linked with the diversification of the mammalian dentition.

Our results also suggest, for the first time, that the evolution of mammalian dental patterns arose through strong positive selection of genes that previously were principally involved in other functions. This is strong evidence that evolution and diversification of teeth arose through modification of genes that had previously been involved in others networks.

Supplementary Material

Supplementary tables S1–S9 and figures S1–S6 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) for financial support to J.P.M. (SFRH/BD/65245/2009) and S.P. (SFRH/BD/47938/2008). S.J.O. was supported in part by Russian Ministry of Science Mega-grant no. 11.G34.31.0068. A.A. was partially supported by the Strategic Funding UID/Multi/04423/2013 through national funds provided by FCT and European Regional Development Fund (ERDF) in the framework of the program PT2020 and the FCT project PTDC/AAG-GLO/6887/2014 (POCI-01-0124-FEDER-016845).

Literature Cited

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol. 55:539–552.

Armfield BA, Zheng Z, Bajpai S, Vinyard CJ, Thewissen J. 2013. Development and evolution of the unique cetacean dentition. PeerJ. 1:e24.

Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.

Barrett T, et al. 2005. NCBI GEO: mining millions of expression profiles–database and tools. Nucleic Acids Res. 33:D562–D566.

Barrett T, et al. 2013. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 41:D991–D995.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 289–300.

Capra JA, Williams AG, Pollard KS. 2012. ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. PLoS Comput Biol. 8:e1002567.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Chatr-Aryamontri A, et al. 2013. The BioGRID interaction database: 2013 update. Nucleic Acids Res. 41:D816–D823.

Cohen BA, Mitra RD, Hughes JD, Church GM. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet. 26:183–186.

Collins MO, Yu L, Campuzano I, Grant SG, Choudhary JS. 2008. Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. Mol Cell Proteomics 7:1331–1348.

Dasmeh P, Serohijos AW, Kepp KP, Shakhnovich EI. 2013. Positively selected sites in cetacean myoglobins contribute to protein stability. PLoS Comput Biol. 9:e1002929.

Davit-Beal T, Tucker AS, Sire JY. 2009. Loss of teeth and enamel in tetrapods: fossil record, genetic data and morphological adaptations. J Anat. 214:477–501.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Flicek P, et al. 2012. Ensembl 2012. Nucleic Acids Res. 40:D84–D90.

Fraser GJ, et al. 2009. An ancient gene network is co-opted for teeth on old and new jaws. PLoS Biol. 7:e31.

Gazave E, Marques-Bonet T, Fernando O, Charlesworth B, Navarro A. 2007. Patterns and rates of intron divergence between humans and chimpanzees. Genome Biol. 8:R21.

Goldman M, et al. 2013. The UCSC Cancer Genomics Browser: update 2013. Nucleic Acids Res. 41:D949–D954.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 27:221–224.

Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 537:113–137.

Hare MP, Palumbi SR. 2003. High intron sequence conservation across three mammalian orders suggests functional constraints. Mol Biol Evol. 20:969–978.

Harris MA, et al. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32:D258–D261.

Henricson A, Forslund K, Sonnhammer EL. 2010. Orthology confers intron position conservation. BMC Genomics 11:412.

Hillson S. 2005. Teeth. University College London, Cambridge: Cambridge University Press.

Horvath JE, et al. 2014. Genetic comparisons yield insight into the evolution of enamel thickness during human evolution. J Hum Evol. 73:75–87.

Houlston RS, et al. 2008. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet. 40:1426–1435.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4:44–57.

Jarvinen E, Tummers M, Thesleff I. 2009. The role of the dental lamina in mammalian tooth replacement. J Exp Zool B Mol Dev Evol. 312B:281–291.

Jernvall J, Thesleff I. 2012. Tooth shape formation and tooth renewal: evolving with the same signals. Development 139:3487–3497.

Jonsson JJ, Foresman MD, Wilson N, McIvor RS. 1992. Intron requirement for expression of the human purine nucleoside phosphorylase gene. Nucleic Acids Res. 20:3191–3198.

Kalmar L, Homola D, Varga G, Tompa P. 2012. Structural disorder in proteins brings order to crystal growth in biomineralization. Bone 51:528–534.

Katzman S, et al. 2007. Human genome ultraconserved elements are ultraselected. Science 317:915.

Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. 16:980–989.

Kelley JL, Swanson WJ. 2008. Dietary change and adaptive evolution of enamelin in humans and among primates. Genetics 178:1595–1603.

Kent WJ, et al. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.

Kim KM, et al. 2012. Gene expression profiling of oral epithelium during tooth development. Arch Oral Biol. 57:1100–1107.

Kin T, Ono Y. 2007. Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat. Bioinformatics 23:2945–2946.
Knight R, et al. 2007. PyCogent: a toolkit for making sense from sequence. Genome Biol. 8:R171.
Koussoulakou DS, Margaritis LH, Koussoulakos SL. 2009. A curriculum vitae of teeth: evolution, generation, regeneration. Int J Biol Sci. 5:226–243.
Lachke SA, et al. 2012. iSyTE: integrated Systems Tool for Eye gene discovery. Invest Ophthalmol Vis Sci. 53:1617–1627.
Landin MA, Shabestari M, Babaie E, Reseland JE, Osmundsen H. 2012. Gene expression profiling during murine tooth development. Front Genet. 3:139.
Laulederkind SJ, et al. 2013. The Rat Genome Database 2013–data, tools and users. Brief Bioinform. 14:520–526.
Lin D, et al. 2007. Expression survey of genes critical for tooth development in the human embryonic tooth germ. Dev Dyn. 236:1307–1312.
Maldonado E, Khan I, Philip S, Vasconcelos V, Antunes A. 2013. EASER: Ensembl Easy Sequence Retriever. Evol Bioinform Online 9:487–490.
Markova-Raina P, Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. Genome Res. 21:863–874.
Mikkola ML. 2009. Controlling the number of tooth rows. Sci Signal. 2:pe53.
Mina M, Kollar EJ. 1987. The induction of odontogenesis in non-dental mesenchyme combined with early murine mandibular arch epithelium. Arch Oral Biol. 32:123–127.
Mitsiadis TA, Graf D. 2009. Cell fate determination during tooth development and regeneration. Birth Defects Res C Embryo Today 87:199–211.
Morgan CC, et al. 2012. Colon cancer associated genes exhibit signatures of positive selection at functionally significant positions. BMC Evol Biol. 12:114.
Nielsen R, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3:e170.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.
Nilsson J, Grahn M, Wright AP. 2011. Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. Genome Biol. 12:R65.
Nylander JAA. 2004. MrAIC.pl. Program distributed by the author. Uppsala: Evolutionary Biology Centre.
Oshima RG, Abrams L, Kulesh D. 1990. Activation of an intron enhancer within the keratin 18 gene by expression of c-fos and c-jun in undifferentiated F9 embryonal carcinoma cells. Genes Dev. 4:835–848.
Pemberton TJ, et al. 2007. Identification of novel genes expressed during mouse tooth development by microarray gene expression analysis. Dev Dyn. 236:2245–2257.
Penn O, Privman E, Landan G, Graur D, Pupko T. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol. 27:1759–1767.
Philip S, et al. 2012. Fish lateral line innovation: insights into the evolutionary genomic dynamics of a unique mechanosensory organ. Mol Biol Evol. 29:3887–3898.
Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20:110–121.
Salazar-Ciudad I, Jernvall J. 2004. How different types of pattern formation mechanisms affect the evolution of form and development. Evol Dev. 6:6–16.
Shimoyama M, et al. 2011. RGD: a comparative genomics platform. Hum Genomics 5:124–129.
Simon-Sanchez J, et al. 2009. Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nat Genet. 41:1308–1312.
Smith MM, Coates MI. 1998. Evolutionary origins of the vertebrate dentition: phylogenetic patterns and developmental evolution. Eur J Oral Sci. 106(Suppl 1):482–500.
Smith SA, Dunn CW. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24:715–716.
Stark C, et al. 2006. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34:D535–D539.
Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci. 100:9440–9445.
Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 56:564–577.
Thesleff I. 2006. The genetic basis of tooth development and dental defects. Am J Med Genet A. 140:2530–2535.
Toll-Riera M, Bostick D, Alba MM, Plotkin JB. 2012. Structure and age jointly influence rates of protein evolution. PLoS Comput Biol. 8:e1002542.
Volponi AA, Pang Y, Sharpe PT. 2010. Stem cell-based biological tooth repair and regeneration. Trends Cell Biol. 20:715–722.
Woo YH, Walker M, Churchill GA. 2010. Coordinated expression domains in mammalian genomes. PLoS One 5:e12158.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.
Yi H, et al. 2010. Gene expression atlas for human embryogenesis. Faseb J. 24:3341–3350.
Zhang T, et al. 2012. SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. J Biomol Struct Dyn. 29:799–813.
Zhang YD, Chen Z, Song YQ, Liu C, Chen YP. 2005. Making a tooth: growth factors, transcription factors, and stem cells. Cell Res. 15:301–316.

Associate editor: Marta Barluenga