Round #1

by Didier Casane, 31 Jan 2023 15:01
Manuscript: **https://doi.org/10.1101/2022.12.09.519446** version 1

This preprint merits a revision

Decision: revise

Three reviewers gave very positive feedback on this manuscript. They also suggest that some changes be made to improve its readability and understanding of the results by most readers. I suggest that the most important ones be done. Many minor concerns should also be considered too.

E1: 1) The authors assembled a dataset of 11 dental genes: nine genes having well-characterized functions and/or expression patterns tied to tooth development and frequently pseudogenized in edentulous and enamelless taxa, and two other genes expressed in teeth but do not have the same patterns of inactivation. I understand this selection that makes sense, but I was wondering if it could be possible to enlarge the set of genes, for example by looking for pseudogenes in xenarthran genomes and checking those involved in teeth development using mammal gene expression databases. I was wondering if these 11 genes represent a small or a large subset of teeth specific genes. If it is a small subsets and if the ancestral branch is small compared to the xenarthran tree, even if some pseudogenes are shared by all xenarthrans, because there are few of them we need a large gene set to find those that appeared in the common ancestor. Thus, no common loss of function mutations in tooth specific genes could be the consequence of a too small number of genes examined and not evidence of normal teeth and gingiva in the common ancestor.

AE1: There are more than 150 genes involved in tooth development but most of them are not tooth specific as they are situated upstream in the development regulation network. We here chose to focus on tooth-structure genes for which previous evidence have shown that they have been pseudogenized in enamelless and/or toothless species and/or in which mutations in human/mouse are linked to diseases/phenotypes. A genomic screen for inactivating mutations in all tooth-related genes would be interesting but is beyond the scope of this paper in which we tried to be exhaustive in species sampling to reconstruct the details of dental gene evolution linked to tooth vestigialization across the xenarthran radiation. We did, however, add a comment at the end of the discussion raising the possibility that newly discovered genes may provide further resolution to this question.

E2: 2) A reviewer wrote: Regarding the methodology, although I understand that they accomplish the goal of getting the sequence of the genes using different approaches, adding a schematic figure showing which method was used for which species would help to understand more easily. Another reviewer wrote: this data collection might also be one of the few limitations of the paper: even with such a broad effort, a lot of genetic information is missing from the final dataset. Except for the few fully sequenced analyzed, it is unclear whether the missing exons of various genes are the reflection of a biological reality or rather of the incomplete molecular sampling due to imperfect amplification or capture and/or lack of depth in the sequencing. This difficulty is exemplified by the DMP1 and MEPE datasets (see below). Since these missing pieces of genes might themselves be involved in the identification of inactivated genes, it slightly blurs the general message. The limited completeness of the dataset for this marker (and the others too) should be directly accessible from the main text to be obvious to help contextualize the interpretation.
Thus, for the 11 genes examined, a clear description of the data completeness is necessary.

AE2: We understand and agree with the basic argument of this comment. We attempted to provide this information in the supplementary tables, but have now realized that someone unfamiliar with

xenarthran phylogenetics would have difficulties discerning how to interpret the relative significance of our missing data. To that end, we have created a new set of supplementary figures, which we believe should adequately display where data are missing and, therefore, support our ability to reconstruct mutational events in early xenarthran history. We acknowledge that we cannot sufficiently describe every pseudogenization event within the xenarthran radiation given missing data. However, given that our research question revolved around dental evolution in the earliest branches of this clade, we believe the visual representations of these data in figures will satisfy the reviewers and specialists interested in this question. The new figures in question are supplementary figures S3–S13. We should add that the sequence alignments are also made available as supplementary datasets for other readers to examine directly.

E3: 3) A reviewer wrote: I also like the evolutionary rationale behind reconstructing dN/dS values and when the inactivation process occurred in the xenarthran phylogeny. I recommend creating a schematic figure showing the rationale. Being more didactic will open the paper to a broader audience.
I agree, although crafting such a figure may not be an easy task.

AE3: We completely understand why this would be of use to readers given the complexity of these analyses, and as such, we have constructed a new supplementary figure (S2) that we believe should address this concern. In short, it shows the workflow used for calculating dN/dS ratios in the various models using an example gene with direct references to the summarized results in the supplementary tables. The hope is that readers can refer to this graphical overview and the corresponding supplementary tables in order to ensure comprehension of our approach.

Reviews

Reviewed by Juan Opazo, 30 Dec 2022 18:32

R1.1: This work by Emerling et al. attempts to unravel the genetic bases of dental regression in a group of mammals, xenarthrans, that present different degrees of reversals. The attractiveness of the system is that exists a good characterization of the genes that are directly involved in the dental phenotype. It represents an outstanding opportunity to disentangle the genetic bases of the dental regression on xenarthrans and show the nature of the evolutionary process. I like how the introduction is written, from general aspects of regressive evolution, a description of the study system (dentition), the goodness of the selected taxonomic group, and the kind of questions that can be answered.

AR1.1: Thank you for the encouraging comment.

R1.2: The taxonomic sampling is well thought out to answer the proposed questions. It includes a diverse sampling for the ingroup and also for the outgroup.

AR1.2: Thank you.

R1.3: Regarding the methodology, although I understand that they accomplish the goal of getting the sequence of the genes using different approaches, adding a schematic figure showing which method was used for which species would help to understand more easily.

AR1.3: To address this comment, we have created a supplementary figure (new Figure S1) showing the phylogenetic distribution of the xenarthran taxa with symbols indicating the methodology used to reconstruct genes in our dataset. This will supplement our supplementary table that already included this information (Table S2), and we hope will make it easier for readers to keep track of which species had genes reconstructed from each method.

R1.4: The dN/dS section needs to be clarified for me. How it is described does not allow the reader to create a mental image of the tree and the estimated omegas. In fact, the authors clarify not to confuse what they are describing with the free ratio model, referring to a paper that is not the best way to do it.

AR1.4: As discussed above, we have addressed this by creating a new supplementary figure (Figure S2), which walks through the analysis of a gene example (AMELX) and the corresponding supplementary table. As such, we believe that this will be much more clear for readers.

R1.5: The result section is ok, where they described the evolutionary events related to the inactivation of genes. Figure 1 is very informative, although it is also very busy. I suggest using common names instead of scientific names. I like figure 2, which shows how the genes got inactivated in the different branches of the xenarthran tree of life. However, I am afraid that the final version will be too small to see the details easily, it is possible to show this information in two figures? I also suggest using common names.

AR1.5: We appreciate the comments to our figures. Responses to each below:

Figure 1: While common names can be useful for fluent English speakers, in our experience they can be confusing for those specialists that are less familiar with the English common names. As such, we prefer to keep their latin binomial names. We do however want to highlight that the common names for the clades (e.g., "long-nosed armadillos", "anteaters") are listed towards the right of the figure, to aid those who are familiar with English common names.

Figure 2: We found this to be a very reasonable and helpful suggestion, and have decided to split this figure into new Figures 2 and 3. The former is dedicated to the mutations of anteaters and sloths, and the latter to those of armadillos.

R1.6: I also like the evolutionary rationale behind reconstructing dN/dS values and when the inactivation process occurred in the xenarthran phylogeny. I recommend creating a schematic figure showing the rationale. Being more didactic will open the paper to a broader audience.

AR1.6: As discussed in AR1.4, we believe our new supplementary figure (Figure S2) will clarify this.

R1.7: Regarding figure 3, why are the authors rooting the tree with afrotherians? Do they think the issue related to the first branching out in placental mammals is "solved"? Did they try the same analyses using atlantogenata as a sister group of boreoeutheria? I am also afraid that the final version will be too small to see the details, it is possible to show this information in two figures?

AR1.7: As for the issue of the placental mammal root, this is solely based on our reference phylogeny (Emerling et al. 2015), as cited in the methods portion of the paper. We did not attempt to repeat the same analyses with alternative rooting, given that in our experience, the changes in dN/dS estimates are negligible given generally small differences in branch length such as the ones induced by a different rooting of the placental tree. To improve the figure legibility, we have split it into three distinct panels (Figure 4A,B,C) that allows readers to see the details more clearly. Moreover, the original outputs of the CoEvol analyses obtained from Drawtree are provided as Supplementary Figures S25-S35.

R1.8: There is a gene called gremlin 2, for which genetic experiments have shown its role in tooth development (Brommage et al., 2014). In fact, it has been demonstrated that gremlin 2 deficient mice have upper and lower incisor teeth with a markedly reduced breadth and depth, where the upper incisors are more severely affected than the lower ones (Vogel et al., 2015). Furthermore, a study in

which the evolutionary history of this gene was examined showed that gremlin2 loss coincides with a lack of upper jaw incisors in ruminants (Opazo et al. 2017). Therefore, it may be a good idea to include it in this study.
Brommage R, et al. 2014. Bone Research. PMID: 26273529.
Vogel P, et al. 2015. Veterinary Pathology. PMID: 24686385.
Opazo JC, et al. 2017. PeerJ. PMID: 28149683.

AR1.8: Thank you for your suggestion regarding the possibility of including GREM2. We did some BLAST searches against available xenarthran assemblies and found that this mono-exonic gene appears to be fully functional in all species examined. Therefore, we did not consider it further in our study.

Reviewed by Régis Debruyne, 17 Jan 2023 09:01

R2.1: The manuscript entitled “Genomic data suggest parallel dental vestigialization within the xenarthran radiation” by C.A. Emerling et al. presents evidence for the parallel evolution and gradual decay in dental regression in xenarthran lineages. It is a very consistent and robust manuscript, based on a well-designed analytical methodology.
The authors ground their hypotheses on two major lines of evidence: sequence analysis of gene pseudogenization and dN:dS analysis of nucleotide sequences, for eleven core genes involved in enamel and tooth formation. Both provide compelling support to the main hypotheses of independent and stepwise loss for enamel/teeth.
The scientific background and the specific aims of the manuscript are well-developed while remaining clear to the non-specialist. The manuscript is well-written and the authors provide a wealth of details about how the dataset was constructed and analyzed. The in-text figures are designed clearly while conveying a large amount of information. All necessary material is presented or made available to support the authors’ claims; relevant supplementary tables (26 six of them, all meaningful to me) and figures also provide extra details to the keen reader. I appreciated that the discussion addresses how the molecular findings integrate with the (scarce) paleontological evidence available from the relevant Eocene/Oligocene Xenarthra fossils, and the changes in the ecology/diet of the sloths

AR2.1: Thank you for your kind comments and constructive feedback.

R2.2: I only regret that some sentences in the main text sometimes lean towards a non-necessary finalist presentation of the evolutionary processes. For example l.34 reads “sloths halted their dental regression; l.468-9 reads “stem dasypodid armadillos independently inactivated AMTN”. My remark might only reflect the fact that English is not my mother tongue, and that this phrasing sounds finalist to the French reader that I am.

AR2.2: We believe this concern to be an issue with using the active voice rather than the passive voice, such that from our writing that it appears as if sloths intended to halt their dental regression and dasypodid armadillos willfully inactivated AMTN. By contrast, the passive voice would convey that these are events that occurred in sloths and armadillos. Assuming our interpretation of this criticism is correct, we have modified the text accordingly:

*“*sloths halted their dental regression” *was changed to* “Next, whereas dental regression appears to have halted in sloths…”

and “stem dasypodid armadillos independently inactivated *AMTN*” *was changed to* ”Our data suggest that *AMTN* and *ODAM* were independently inactivated in stem chlamyphorid and stem dasypodid armadillos…”

We hope this addresses your concern.

R2.3: I am fully convinced by the various claims made in the manuscript concerning the independence, parallelism, and gradual characteristics of the dental regression processes identified in the various subclades within Xenarthra. I agree with the authors that the lack of SIM signal observed in the DMP1 and MEPE genes only reinforces the observations made on the 9 other genes more specifically involved in enamel/tooth development.
I am a bit surprised that not a single SIM was recovered for any of the genes under scrutiny for the ancestral branch of the entire Xenarthra super-order, considering that dental regression is a feature observed even in the earliest-known xenarthran fossils (as indeed the authors acknowledge in the discussion). I think that this observation might deserve a little more consideration in the discussion: it could be emphasized that this is a clear limitation of such a functional study based solely on the analysis of coding sequences and that the processes identified here might just reflect one side of the coin. Adding to this, I wonder why only the coding sequences of the focal genes were targeted. It is obvious from the results (AMTN, ODAM) that key mutations leading to pseudogenization might have occurred in non-coding sequences of these genes or promoter regions also. Were the intron sequences excluded for any specific reason (like strong divergence among xenarthrans, or else)? I would appreciate the authors briefly justify this choice.

AR2.3: We agree with the points brought up here, namely that using protein coding regions alone is a limitation in reconstructing the dental history of xenarthrans. Indeed, this was part of the rationale behind performing dN/dS analyses, as described in the results:

"Given that shared inactivating mutations provide only a minimum probable date for inactivation, they may underestimate the timing of the onset of relaxed selection on a gene. Gene dysfunction may predate the fixation of a frameshift indel or premature stop codon, e.g., due to disruption of non-coding elements."

Furthermore, we alluded to this possibility in the discussion:

"It remains possible that non-coding mutations leading to dental regression could have accumulated prior to the mutations more typically characteristic of pseudogenes, at least by the origin of cingulates."

As far as non-coding regions are concerned, while we are certainly interested in the possibility of reconstructing ancestral genes in xenarthrans, including promoters, enhancers, introns, etc., this is beyond the scope of this study. Pseudogenes provide a relatively concrete link of genetics to phenotype, given that they naturally represent the absence of gene function and therefore protein product. By contrast, analyzing non-coding elements in this context would seem to necessitate both (1) conserved elements, which don't always exist, particularly when pseudogenization has occurred at these time scales, and (2) precise functional data that can tie specific mutational data to a phenotypic consequence (e.g., downregulating genes, change tissue expression of genes, etc…). We hope to be able to explore this phenomenon further, including with Xenarthra, in future projects.

*To address this valid criticism specifically, we have added some discussion to encourage future work on this system: "*Understanding of this system would benefit greatly from analyzing non-coding elements and functional data, given that mutations outside of the protein-coding regions of these genes may pre-date frameshift indels, premature stop codons and similar inactivating mutations."

R2.4: Gathering the gene (exonic) sequences dataset analyzed in the paper for such an extensive sample within modern Xenarthrans (31 species represented) is quite impressive. The sources of the data analyzed are extremely diverse: whole-genome sequencing, shotgun and target capture sequencing, targeted amplicon sequencing. Yet, this data collection might also be one of the few

limitations of the paper: even with such a broad effort, a lot of genetic information is missing from the final dataset. Except for the few fully sequenced analyzed, it is unclear whether the missing exons of various genes are the reflection of a biological reality or rather of the incomplete molecular sampling due to imperfect amplification or capture and/or lack of depth in the sequencing. This difficulty is exemplified by the DMP1 and MEPE datasets (see below). Since these missing pieces of genes might themselves be involved in the identification of inactivated genes, it slightly blurs the general message.

AR2.4: This is an understandable criticism, as it was something we were keenly aware of while compiling this dataset. However, we believe that it does not undercut our conclusions in a way that is ultimately meaningful. The orthologs for which we have missing data prevent us from identifying every significant mutational event in the protein-coding regions of these genes. For example, we cannot say if a particular stop codon is shared among all Cabassous *species or just a few. As such, for intrageneric diversification events, our ability to describe the mutational history of these genes is diminished. However, we believe our dataset is complete enough to allow us to make inferences about the deep history of Xenarthra, and particularly pertaining to the stem Xenarthra, stem Pilosa, stem Cingulata, stem Vermilingua, stem Folivora, stem Dasypodidae, and stem Chlamyphoridae branches. This is due to the fact that we were able to work with whole genome assemblies and/or sequencing from major lineages of xenarthrans: both sloth genera (Choloepus [2 spp.], Bradypus), all three anteater genera (Tamandua, Cyclopes, Myrmecophaga), the sole dasypodid genus (Dasypus), and species from each chlamyphorid subfamily: Euphractinae (Chaetophractus), Tolypeutinae (Cabassous, Tolypeutes) and Chlamyphorinae (Chlamyphorus, Calyptophractus). For the stem Dasypodidae branch, the whole genome assembly of Dasypus novemcinctus was sufficiently supplemented by exon capture, PCR and (in one case) whole genome sequencing data.*

*We believe that the sufficiency of our data for the purposes of our conclusions will be further underscored by the inclusion of the supplementary figures (S3–S13) discussed above. We have added a sentence to our results to further clarify this point: "*While we were unable to recover the complete coding sequence for every gene in every species, the phylogenetic distribution of taxa derived from whole genome sequencing means that this inference is unlikely to be the result of missing data (Supplementary Figures S3-S13)."

R2.5: For the ACPT gene, specifically, the authors state that its CDSs were not included in the baits designed for target sequence. Why? Due to this choice, a majority (19 out of 31) of xenarthran taxa are not directly represented in the ACPT alignment. Thus direct evidence of pseudogenization for this gene is scarce. However, figure 1 indicates that for 11 non-sequenced armadillo species, the dN/dS ratio estimates suggest a possible gene inactivation. Could the authors be more specific as to how this result is obtained since only a single direct sequence is available for all 12 most basal armadillos?

AR2.5: Regarding the relatively minimal sequences for ACPT (now referred to as ACP4), this is merely a consequence of the timing of when the function of ACP4 was characterized as being associated with amelogenesis imperfecta versus when the exon capture baits were designed. In short, by the timing the baits were designed and by the time the exon capture was completed, the study in which ACP4's function was described was only just published (Seymen et al. 2016). As such, we were only able to obtain this gene for species for which we had whole genome sequences. We have added a detail about this to the methods to explain this discrepancy.

However, we have since performed whole genome sequencing for additional xenarthran genomes, and better assemblies have become available for a few additional species, allowing us to fill in the gaps a bit and provide further resolution to this gene's history in Xenarthra.

The species we were able to add include a tolypeutine armadillo (Priodontes maximus), a euphractine armadillo (Euphractus sexcinctus), a dasypodid armadillo (Dasypus kappleri), a three-fingered sloth (Bradypus tridactylus), plus some improved assemblies for a few species for which we already had sequence data. Furthermore, the addition of these sequences allowed for more refined BLAST searches and improved alignments, resulting in an even more complete ACP4 dataset. We then reran all of our analyses for this gene and updated the results for ACP4 across the manuscript (methods, results, discussion), figures, supplementary tables, supplementary figures and supplementary datasets.

The highlights of these new results are as follows:

1. *We found four shared inactivating mutations (SIMs) across pilosans, suggesting this gene was inactivated in a stem pilosan. This pushes the pseudogenization date earlier than what we previously reported (stem Vermilingua, stem Folivora).*
2. *Thanks to the addition of* Euphractus sexcinctus*, we found two SIMs shared across all chlamyphorids, suggesting this gene was inactivated in a stem chlamyphorid. This was a plausible possibility suggested by dN/dS ratio analyses, but now with confirmation.*
3. *With the addition of* Dasypus kappleri*, we found 13 SIMs that likely accumulated on the stem Dasypodidae branch. This was a likely result based on dN/dS analyses, and the sheer volume of inactivating mutations observed in* Dasypus novemcinctus*, though again this provided confirmation of that prediction.*
4. *We found our first piece of evidence for an elevated dN/dS ratio in ACP4 on the stem Xenarthra branch, though this model was only marginally significant compared to a model where the branch was fixed with the background ($p = 0.049$). However, we believe that this is not suggestive of pseudogenization on the stem xenarthran branch given (1) no evidence of SIMs across all armadillos and (2) the stem Cingulata (armadillo) branch is suggestive of purifying selection: $w = 0.1$, background = 0.18, being significantly different from a model in which stem Cingulata branch is fixed at 1 ($p = 0.00004$).*
5. *The inactivation of ACP4 specifically is associated with hypoplastic amelogenesis imperfecta in humans and animal models, resulting in thin enamel. This leads us to the interpretation that enamel became thin on the stem Dasypodidae and stem Chlamyphoridae branches, consistent with evidence in fossil armadillos (Astegotherium, Utaetus)*

Given the new ACP4 results, the dN/dS comment about figure 1 is now moot, and that information has been removed from the figure.

Please note that while our additional whole genome shotgun data would allow us to add a bit of data to some of the other genes, we have every reason to believe that these additions would be largely insignificant and therefore prove to be an unnecessary volume of work with no meaningful benefit to our study. As such, we only added data for ACP4, given the large gaps in our dataset for this gene.

R2.6: When compared with the information conveyed in figure 1, it seems that some of the identifications for the genes between 'missing-or-pseudogene_phylogeny-based/delta', 'pseudogenized/psi', 'unknown/?' and 'pseudogene_dN:dS-based/omega' lack consistency among the various markers and taxa. For example, some 'delta' identifications are made for some taxa for which not a single exon sequence of the marker of interest was retrieved (as 'missing'), but sometimes it is given for taxa that show only a few of the marker exons (as pseudogene based on the phylogeny'). I would rather have these two situations split into two different summary letters, for they do not convey the same information at all.

AR2.6: Perhaps our meaning of what delta signifies was not clear. Delta indicates instances where a gene is inferred to be a pseudogene based on the phylogenetic distribution of inactivating mutations, but the relevant data are missing in that particular species. Those missing data might be due to the complete absence of the gene at one extreme, or just the absence of the relevant site (e.g., missing exon, part of an exon, etc.). For instance, Cabassous tatouay *was given a delta symbol for AMTN and ODAM. For AMTN,* C. tatouay *had no data, but for ODAM we recovered only exon 9. In both cases, there was no positive evidence of pseudogenization, but the phylogenetic distribution of shared mutations suggests that both AMTN and ODAM are indeed pseudogenes in* C. tatouay.

We modified the figure caption such that we hope it is further clarified:

*"*Δ = no positive evidence of pseudogenization, but gene inactivation or deletion inferred from phylogenetic distribution of shared mutations"

Please note that we have removed the dN/dS / omega symbol as discussed above. The full set of categories is now listed as follows:

*"*Ψ = positive evidence of pseudogenization; Δ = no positive evidence of pseudogenization, but gene inactivation or deletion inferred from phylogenetic distribution of shared mutations; ? = no data; empty box = gene putatively intact."

R2.7: I'll provide a few examples, below, of situations that put the emphasis on the not-so-obvious link between the supplementary tables S5 to S15 and the interpretation that is made of them within figure 1.
Table S6 – AMBN gene. The sequence for Cabassous tatouay is represented in the dataset by a single exon (exon6) which is deemed putatively functional, but the other 9 expected exons are missing. This sequence shows a "?" in figure 1. However, nine other taxa lack one or several exons (up to 5 for Euphractus sexcinctus) for this gene and yet are all deemed functional for AMBN, because no obvious signature of pseudogenization is found in the remaining exons. Isn't there a risk, in such a situation – which corresponds to the majority of the markers analyzed – to underestimate the actual inactivation of the markers? In other words, how can we be confident that, with one or several missing exons, the protein produced is still efficient and not sub-functional? And from how many absent exons should one deem that the protein is not functional anymore? It seems fairly arbitrary to me at this stage (and if so, which is fine, it should at least be specified).

AR2.7: We understand and think this is a valid criticism in regards to the putative functionality of a particular gene. At most, we can provide positive evidence of gene inactivation (frameshift indels, premature stop codons, etc…) or strongly infer pseudogenization based on the phylogenetic distribution of shared inactivating mutations (e.g., all of the anteaters and sloths for which we have data share a frameshift indel, therefore all pilosans likely have a dysfunctional gene). However, as discussed above and in the manuscript, we may certainly be underestimating the distribution of pseudogenes.

First, the goal of this project is not to describe the functionality of all 11 genes for every single species. Instead, we are trying to test whether the loss of genes is suggestive of dental regression early in xenarthran history (e.g., stem Xenarthra, stem Cingulata, stem Pilosa) or relatively more recent gene losses (e.g., stem Cabassous*, stem Euphractinae, etc…). In such a case, while having complete sequences would be ideal, it is not necessary. We simply needed to ensure that we had enough sequences across the xenarthran phylogeny to estimate early gene losses, which is why having whole genomes was critical to achieve this goal (given that PCR and exon capture are less efficient with eroded pseudogenes). We believe we have more than adequate phylogenetic and genetic coverage, and that this will be more discernible to readers with the addition of Supplementary Figures S3-S13.*

Second, the incompleteness of genes in some species in turn provides the rationale for performing dN/dS ratio analyses. Given that relaxation of selection on any particular gene may predate positive evidence of pseudogenization, we tested for evidence of relaxed selection on the early xenarthran branches. While we found nothing to suggest this (with the possible exception of ACP4, discussed above), as you have pointed out, and as suggested by the fossil record, there could well have been genomic erosion of tooth genes, albeit we hypothesize that such a signature may be contained in non-coding elements.

*That said, upon reading your comment, it became apparent to us that the use of the question mark (?) to signify "*unknown*" in the figure was not particularly informative. To make better use of this symbol, we have decided to only employ it for instances where we did not recover the gene at all, and the gene is not suggested to be a pseudogene based on phylogenetic bracketing. We also clarified for genes indicated with an empty box that the gene is "putatively" functional, allowing for the possibility of undiscovered mutations, especially in cases with a lot of missing data.*

Relevant portions of the caption below:
"? = no data; empty box = gene putatively intact."

R2.8: Table S7 – AMELX gene. For Cabassous tatouay, again, only one exon out of 4 is present in the dataset (and not mutated). Yet this time the gene is deemed inactivated via 'delta' in figure 1. Why is the interpretation different for this marker from the AMBN case?

AR2.8: Discussed above.

R2.9: Table S8 – AMTN gene. Priodontes maximus shows only one (non-mutated) exon out of the 7 expected and is identified with a delta (pseudogenized based on phylogeny I guess) like Totypeutes tricinctus for which not a single exon was found (thus recorded as a missing gene for this one). Again the use of a delta for both categories seems misleading to me.

AR2.9: Discussed above.

R2.10: Table S14 – ODAM mutations. Among Vermilingua, Myrmecophaga tridactyla show the 10 expected exons of the marker, with 6 of them showing inactivating mutations, and yet it is identified as a "delta" - inactivated based on the phylogenetic bracketing – instead of the "psi" like Tamandua tetradactyla. Why?

AR2.10: This was an error and should have shown the psi symbol, given that we found positive evidence of pseudogenization. Thank you very much for your thoroughness in checking these minutiae and pointing them out to us.

R2.11: Table S9 – DMP1. All 31 taxa are presented as having a functional DMP1 in figure 1. Yet, the 5 expected exons are only documented for the 11 complete genomes analyzed, whereas the other 20 taxa showed the lack of at least 1 and up to 4 of these exons. Based on the observation that the present exons never show a SIM signature, it sounds reasonable to extrapolate that these sequences are functional as the authors do. Yet, the limited completeness of the dataset for this marker (and the others too) should be directly accessible from the main text to be obvious to help contextualize the interpretation made.

AR2.11: As discussed above, we have added supplementary figures (Figure S7 for DMP1) to aid in the interpretation of our results based on the completeness of our dataset. Note that these figures show the proportional sizes of the exons. In this case, it allows readers to see that we recovered the largest exon in DMP1 (exon 5) in all species, which is much bigger than the remaining four exons. As such, the functionality of this gene can be all but assured across the various taxa, a likelihood supported by the dN/dS ratio analyses.

Reviewed by Nicolas Pollet, 30 Dec 2022 13:33

Dear Christopher Emerling and colleagues,

R3.1: I read with interest your preprint entitled « Genomic data suggest parallel dental vestigialization within the xenarthran radiation ». In this research article, you report your findings on the patterns of sequence evolution related to the vestigialization of teeth development in xenarthran, a group of mammals known to have experienced changes in their teeth ranging from regressive modifications to a complete absence.
You collected data and explored the evolution of sequences corresponding to eleven known genes involved in tooth development in anteaters, armadillos and slugs. You present a phylogenomic analysis that revealed different patterns of mutations across Xenarthra. You conclude that the regressive evolution linked to tooth development in these mammals could not be linked to a single mutation, and instead followed parallel trajectories in the different clades and occurred relatively rapidly.
I find that the title reflects the main message of the study.
I find that the abstract is relatively concise and presents the main results and conclusions of the study.
In the introduction, the theme of “regressive” evolution is presented as the underlying theme of the paper. The specific domain of teeth loss in jawed vertebrates is also explained, along with background information on specific and well-known genes involved in tooth development and differentiation. The Xenarthra taxon and its teeth traits are presented in enough detail so that the reader can understand the research questions and hypothesis. The references provided in the introduction are relevant and cover not only the most recent research but also covers older literature.

AR3.1: We thank you for your constructive comments and positive feedback.

R3.2: In the Materials and methods, the information related to the gene set being analysed is presented in sufficient details. I just checked the gene symbols and became aware that the approved gene symbol for ACPT is now ACP4 (and at the same time I learned that its whole name is testicular acid phosphatase ! https://www.genenames.org/data/gene-symbol-report/#!/hgnc_id/HGNC:14376).

AR3.2: Thank you for noting that. Our impression was that the trend was towards naming it ACPT, but it seems that recent publications are indeed using ACP4. We have changed this throughout the manuscript and supplementary materials.

R3.3: The taxonomic sampling is presented with details on each sample. Protocols for the experiments of molecular genetics are given in a manner enabling reproducibility. The sequences OP966064-OP966335 were not unavailable at the time of review.

AR3.3: The sequences are now publicly available.

R3.4: The procedure for sequence analysis are missing some minor details (e.g. BLAST version).

AR3.4: We have added this detail. Quoted below:

*"...*against the WGS database using discontiguous megablast (BLAST+ 2.8.0)"

R3.5: The Bioproject PRJNA907496 was unavailable at the time of review. I have to say that I releasing the southern naked-tailed armadillo genome assembly as a Zenodo dataset is not acceptable, the international nucleotide sequence database collection (INSDC www.insdc.org) lists the sole repositories for such data.

AR3.5: We have submitted the Cabassous genome Discovar assembly to GenBank and linked it to the Bioproject. Both raw reads (PRJNA907496) and genome assembly (JAQZBX000000000) are now publicly available.

R3.6: In the paragraph Dataset Assembly, line 225 Supplementary dataset S1 does not correspond to sequence alignments, I guess you meant dataset S2-S23. In this paragraph, I was expecting to find the rationale to consider an exon as missing (ie yellow in supp tables S5-S15).

*AR3.6: Regarding the supplementary dataset numbering, thank you for catching this. There was a change in numbering as we uploaded the files to Zenodo. The Zenodo repository has been updated with a new DOI and should contain all supplementary files correctly numbered. As for the missing exon rationale, this has been added under the section entitled "Inactivating mutations and dN/dS ratio analyses". The added sentences are as follows: "*We also noted examples in which exons were not recovered, some of which may be whole exon deletions, though validation would require contiguous genome assemblies."

R3.7: Line 219 you mention that "sequences were assembled and occasionally combined", how does this link to the supp table 2 ? Did you use a cut-off for base quality before proceeding to mutation analysis and did you look for hetero or homozygosity in the case of inactivating mutations?

AR3.7: For the question about line 219, Supplementary Table S2 is summarizing the genes and sources from which they are derived. So for example, Chaetophractus villosus *AMELX is derived from a combination of PCR and exon capture data. In terms of inactivating mutations, as stated in the methods: "*Consensus sequences of mapped reads were called using the 50% majority rule for each targeted gene"*. We generally did not look for heterozygosity for inactivating mutations, though we noted a few such examples that we found (start codon* Chlamyphorus *AMBN; 2-bp deletion* Cyclopes *AMTN; stop* Dasypus hybridus *ENAM; stop* Dasypus sabanicola *MMP20; splice donor mutation* Chlamyphorus *ODAM; stop* Cabassous centralis *ODAM). The focus of our study, however, was not whether any particular mutations were fixed in a species, but rather whether there were shared mutations suggesting gene loss in early branches of Xenarthra. Presumably the vast majority, if not all, such mutations are homozygous, though we expect that some recent mutations may not be fixed yet in certain species.*

R3.8: Overall there is no specific code or script provided. This latter point could be improved as the analysis of the branch model dN/dS ratio with the various models and their statistical analysis (lines 245-266) is of methodological interest.

AR3.8: We are now providing the commands used for the Coevol analyses. For running codeml in PAML, no script can be provided, as this is implemented in the form of control files. Every control file needs to be modified slightly for the particular details, and as such, there are dozens. We believe the parameters for our analyses, however, are clearly articulated in the methods and should be clear to individuals familiar with codeml in PAML especially given the new explanatory supplemental figure (Figure S2).

R3.9: In the results, you reported the findings on gene mutation patterns across xenarthrans in a condensed format in Figure 1. While this figure is extremely informative, I think that it holds too many pieces of information. I have a suggestion to help the reader follow the figure along with the text: you could position the matrix of shared inactivation on the left as it is the starting point in terms of data, and then have the species tree (with the root oriented toward the right). In my opinion, the mix of colors and symbols is extremely difficult to analyze. A possibility would be to split the figure in two, with one panel showing the data matrix, and another panel with the tree and traits. Also, the use of serif characters in figures is usually avoided (you used it for sub-order names and common names).

AR3.9: We thank you for this thoughtful comment on Figure 1. After discussing, we have decided to keep it largely as is given that, while there is a lot of information conveyed, the key points have been summarized: the relative extent and timing of gene losses, highlighting of key predictions in dental morphology with text, and readers can examine the symbols paired with Figures 2 and 3 (formerly just Figure 2) to look at notable gene inactivations in more detail. We have taken your suggestion for changing the relevant text in this figure to sans serif.

R3.10: Line 410-414: I suggest that you follow the scientific convention on significant figures for omega values.

AR3.10: We thank you for this suggestion, but we are unsure what the scientific convention for omega values would be, as this is not something we have ever come across in our readings. Providing two figures after the decimal point, to us, seemed very reasonable in terms of precision and informativeness. If readers want to see the full output of digits for omega values, they are presented in the supplementary tables.

R3.11: In Figure 2, which provides key data, I notice that the alignments of the different genes do not always include the translation, you could maybe insert a consensus translation over the alignments in these cases. The presentation of the different genes does not follow a simple rule (such as alphabetic ordering), so different mutations of the same gene are not presented side by side. Since the objective here is to show shared inactive mutations, It may be better to include all taxa for a given gene with different SIMs : e.g. AMTN exon1 next to exon4, with all taxa. And to highlight mutations, the use of a dot for identical residues and of lowercase for silent mutations may help.

AR3.11: The translations are only provided when a premature stop codon is present. In other cases, they aren't particularly relevant, as we are simply aiming to summarize the mutations themselves in this figure. As such, we do not believe a consensus translation would be of utility here. As for the genes presented side by side, though we do see the value of this, we have broken up this figure into two (Figures 2 and 3) per the suggestion of another reviewer, and we further believe the key is to highlight the taxonomic distribution of gene losses in the context of early xenarthran history. As such, grouping together stem pilosan mutation, stem vermilinguan mutations, etc… to us is more important to highlight than grouping together specific genes.

R3.12: On figure 3, the color scale is too small to be readable. In the text, you mention lines 423-427 that patterns of relaxed selection are strikingly similar among genes having the same or similar role during tooth development. This statement lacks some explicit description, as a matter of fact, I found that DSPP and MMP20 share a similar pattern, and that EDMA and ODAM also share a similar pattern. You could expand this section to draw conclusions that are more in line with the results.

AR3.12: The figure (now Figure 4) has been spit in three panels to improve readability including the color scale. We also added information on the location of shared inactivating mutations (SIMs) on the different gene trees. The original outputs of the Coevol analyses obtained from Drawtree and used to

make this figure are also provided as Supplementary Figures S25-S35 so that readers can have all details. Concerning your remark on the lack of explicit description supporting the fact that inferred patterns of relaxed selection are similar among genes having similar roles in tooth development, we do not see in what DSPP and MMP20 share a similar pattern of dN/dS variation and we also do not understand to which gene you are referring to as EDMA. We nevertheless expanded the description of the Coevol analysis results reported in Figure 4 to make the comparisons among functional categories more explicit and to underline the usefulness of this approach to illustrate shifts in selection pressures across the phylogeny.

Discussion:

R3.13: Line 442 : I was wondering about the odds to identify shared inactivation mutations across the xenarthrans radiation estimated at 68 million years ago? Intuitively and maybe naively, I would expect that a single ancestral inactivating mutation could be followed by subsequent mutations of different scopes, including larger deletions. The finding that exon 11 of ACPT is deleted in most xenarthran would fit such a scenario.

AR3.13: We thank you for your thoughts on this. Indeed, this is something that can and does happen, both in this and other systems. To better summarize this visually, we used the data summarized in supplementary tables S5-S15 to augment supplementary figures S14-S24. Regarding the odds of detecting mutations, it's a complicated issue that depends on the strength of selection, the size of the gene, the completeness of the gene, etc. As discussed in a comment above, we have added data for ACP4, and now have recovered exon 11 in multiple species.

R3.14: Line 458 a typo at coeval ?

AR3.14: We did check, and "coeval" is indeed the correct word and spelling. To clarify, this is not to mean "Coevol" the program, but a word indicating "having the same age or date of origin; contemporary".

R3.15: In this discussion, I was expecting you to provide some more elements on the limitations of your work and your approach to finding mutations.

AR3.15: We agree, and have added further discussion concerning this. See annotated version of the manuscript.

R3.15: Overall your interpretation of the data takes into account the older and newer studies in the field, and they are reasonable given the provided results. I especially appreciated the broader implications regarding gene loss and evolution, their tempo, and the relationship with life traits.

AR3.15: Thank you.

Supplementary data:

R3.16: I was surprised to find the legends for the supplementary datasets and figures in the supplementary_Tables_S1-S26. I suggest you to add a text file for all supplementary datasets and figures caption, but since it is on the Zenodo page this may be unessential. Dataset S1 does not contain all FASTA alignments, I guess you omitted the range: Dataset S1-S24 !

AR3.16: We added a README.docx file to the Zenodo repository with all supplementary datasets and associated captions. Dataset S1 is the set of baits used in the capture experiments. All other sequence datasets are now provided as individual alignment fasta files (Datasets S2-S23) on Zenodo.

R3.17: The main strength of your study is that I find it extremely well performed, and associated with relevant conclusions.

AR3.17: Thank you for this positive appreciation of our work.

R3.18: The main weakness is that I could not find in the manuscript the details to ensure that all necessary permits or approvals were obtained for collecting the animals. I acknowledge that this can be difficult for Museum samples, but such details should be available for samples from the Animal Tissue Collection of ISEM, for which you gave voucher information. For these vouchers, I could no find any information on the permits that enabled animal collection, even from the cited references. The reference Gibb et al 2016 mention only two permits for Bradypus torquatus and Tolypeutes tricinctus. Moreover, some samples presented here are not described in Gibb et al 2016 (e.g T-1476, T-1722, T-1631, T-2977, T-JL556). Delsuc et al 2018 refers only to Mylodon darwinii. Whether we like it or not, this is an ethical requirement for a scientific publication. In my quality of reviewer, I should also point out that a Benefit-sharing statement is a condition for publication in numerous journals, such as Molecular Ecology (https://onlinelibrary.wiley.com/page/journal/1365294x/homepage/forauthors.html).

AR3.18: Details of all biological samples used in this study and associated museum voucher specimens (when available) are provided in Supplementary Table S1. The Animal Tissue Collection of the ISEM is a registered collection of mainly mammalian tissue samples preserved in ethanol and available for scientific research, which was under the curatorship of the late François Cazeflis since 1986. All xenarthran tissue samples used in this study were collected and preserved in the collection between 1994 and 2008. These are tissues (mainly ear biopsies) that are not linked to proper voucher specimens and were collected in accordance with the relevant legislation on collection and export permits at the time of sampling. These samples were collected prior to the application of the Nagoya Protocol, which implemented the benefit-sharing policy in France in October 2014. However, as advised, we have included a benefit-sharing statement in the revised version to emphasize our willingness to collaborate and to recognize the essential contribution of sample collectors over the years, particularly in French Guiana, where most of the samples in the ISEM collection originate, and where we have developed a long-term collaboration with the Pasteur Institute and the KWATA NGO. These institutions host the JAGUARS tissue collection (http://kwata.net/la-collection-jaguars-pour-l-etude-de-la-biodiversite.html) under the curatorship of Benoit de Thoisy in Cayenne, with whom we regularly exchange and share mammalian tissue samples through formal material transfer agreements as both our institutions are CITES registered (ISEM: FR 34B; KWATA: FR 973A).

We are now providing the sample collection dates in Table S1 and added the following statement in the Materials and Methods section of the paper: "In accordance with the policy of sharing benefits and advantages (APA; TREL1916196S/224), biological material from French Guiana collected after October 2014 has been registered in the JAGUARS collection supported by Kwata NGO, Institut Pasteur de la Guyane, DEAL Guyane, and Collectivité Territoriale de la Guyane. Biological samples from the JAGUARS collection were exchanged through formal material transfer agreements granted by DEAL Guyane". *In addition, as in all previous papers using the ISEM collection, the many people who helped with the sampling are formally named and thanked in the Acknowledgements section.*

We hope this answers any doubts you may have had about our willingness to comply with the legal and ethical requirements for the use of biological resources in scientific publications.