Peer Review File

Manuscript Title: Testosterone histories from tusks reveal woolly mammoth musth episodes

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

This is a really excellent paper that would seem genuinely to open a new window on the biomolecular analysis of fossil or sub-fossil animals. As far as I am aware the analysis of steroid hormones in such materials using robust chromatography-mass spectrometry approaches is highly novel. I especially liked the strenuous lengths the authors have gone to test their approaches and ideas.

The progression from a modern elephant to fossil specimen is exactly the way to develop this type of work, rather than jumping straight into fossil specimens, as is commonly done. I also liked the analysis of the female elephant to lend further credence to the male reproductive patterns that had been revealed in the elephant and mammoth specimens.

Some caution is required in proposing the use of steroids as a sexing tool as this may be confounded by diagenesis. Perhaps this requires more careful testing or at least a caveat added as indicated on the annotated manuscript.

The amount of thought and effort that has gone into this paper is enormous and exquisitely documented in the extremely well-written manuscript. The illustrations do an amazing job at giving the paper life.

I know I'm not supposed annotate the manuscript but I have and have returned this to the editor. I am happy if this is shared with the authors. The comments I have made are entirely positive and only intended congratulate the authors on their exceptional work or raise helpful suggestions for improving the paper. There are only one or two minor technical corrections that are required. The authors have been unusually upfront about the analytical challenges faced in this sort of work and identified areas for improvement. Their data appear internal consistent and robust. A minor query I have is the structures of low mass ions detected in the MRM experiments following CID? Could this detail be added - I failed to find it in the literature I explored?

I have not commented on the endocrinological or ecological aspects of the paper has this is not my field. My comments focus on the fossil and analytical aspects.

This is the best paper in the field biomolecular paleontology or archaeology I have refereed for Nature or any other journal. I recommend publication of this manuscript.

Referee #2 (Remarks to the Author):

The authors present the first evidence of musth in a woolly mammoth by examining testosterone patterns in ancient tusks. Specifically, the authors measured testosterone, progesterone, cortisol, and androstenedione in dentin from an adult male African elephant, an adult male woolly mammoth, and an adult female woolly mammoth. Testosterone patterns in male tusks followed a pattern that would be expected if an animal was experiencing an annual musth. Additionally, testosterone peaks were absent in the female tusk and in a portion of the tusk representing pre-musth years of the male mammoth, adding strength to the findings. The work presented here is the first successful biological validation of dentin hormones and is not only valuable for studying historical specimens but also for assessing endocrine patterns in tusks from modern species.

I applaud the authors for their work; however, I have comments that should be addressed before publication. In particular, the authors need to clarify the sample masses that were extracted for hormone analysis. 30 mg is mentioned in the paper; however, the sample masses listed in the Extended Data section vary greatly. Recent research has shown that sample mass can impact apparent hormone concentrations (Ajó et al. 2022). If the sample masses listed in the Extended Data section are correct, I suggest that the authors conduct an additional analysis to ensure that the mass of the samples does not impact their results. This could be done by weighing out the same sample (or a pooled sample) in range of masses, extracting them using the protocol provided in the paper, and determining whether sample mass impacts their results.

The progesterone and cortisol data included in this paper are interesting; however, the results are only briefly discussed in the Results and Extended Data section and I question whether they add value to the question of musth in woolly mammoths. I suggest that the authors either include additional information on how these hormones are relevant to musth or potentially remove them from the study, if no relevance can be determined.

Additionally, caution should be taken when interpreting the main findings of this study. Although the authors have shown that testosterone measurements in dentin are biologically relevant, as testosterone was shown to peak annually, which is indicative of an annual musth; biological validations for progesterone, cortisol, and androstenedione were not included. Therefore, the authors cannot conclude that they are biologically relevant.

Tremendous job and I am looking forward to the final version of this manuscript.

Detailed comments below:

ABSTRACT

Line 27: More appropriate references for this sentence:

McCormick, S.D. and Romero, L.M., 2017. Conservation endocrinology. BioScience, 67(5), pp.429-442.

Wikelski, M. and Cooke, S.J., 2006. Conservation physiology. Trends in Ecology & Evolution, 21(1), pp.38-46.

Line 28: It is still debated whether steroids in hair reflect historic endocrine patterns (see paper below). I suggest using baleen as an example of a matrix that can provide historic endocrine data.

Kalliokoski, O., Jellestad, F.K. and Murison, R., 2019. A systematic review of studies utilizing hair glucocorticoids as a measure of stress suggests the marker is more appropriate for quantifying short-term stressors. Scientific reports, 9(1), pp.1-14.

INTRODUCTION
Line 53: Include the scientific name of Asian elephants.
Lines 57-59: Provide reference.
Lines 67-71: I suggest including the rationale for analyzing the other hormones presented in this paper here.

RESULTS
Lines 86-87: The scientific name for African elephants and woolly mammoths were provided in the introduction; you can remove them from this section.
Lines 107-110: Are the ranges for the testosterone and androstenedione ratios in dentin provided in the supplementary text?
Lines 104-106: Need references for the effects of oxidation and bacterial degradation of tusks.
Lines 113-114: I suggest removing this sentence as the rest of the paragraph discusses how sexing an individual based on their hormone measurements would be difficult.
Line 125: (Table 1) The female African elephant is noted to have a median testosterone concentration of 0 in Table 1. Can the authors elaborate as to why they think they got this result? Was the concentration simply lower than the LLOQ?
Line 133: The scientific name for African elephants has already been mentioned – please remove.
Line 137: I suggest adding a brief sentence describing the rationale for analysing oxygen isotopes.
Lines 137-139: Reference "Fig. 1" here.
Line 145: I do not see a CT scan of the complete specimen in the supplementary section. Only the CT scans of the 10 individual sections were provided. Were the authors able to count all the growth layers in the tusk using the CT data and confirm that the core taken from the tusk truly represents the last 9 years of the animal's life?
Line 151: I'm curious as to why the authors did not core a section of the adult male elephant tusk at the distal end (presumably pre-musth years) to replicate the work on the male woolly mammoth tusk?
Line 157-159: Link this sentence to the oxygen analysis results.
Line 194: Scientific name already provided.
Line 258: Please elaborate on the features that were used to correlate growth layers between each slab.
Line 265: (Figure 2) Please define the cross symbol in the figure captions for all figures.
Line 266: The heading title for the female woolly mammoth is missing. The male elephant has a heading title on line 131 and the male woolly mammoth has a heading title on line 189.
Lines 267-268: Add "female" to this sentence.
Line 277: Change "was" to "were".
Line 284: (Figure 3) Could the authors adjust the figure so that the pattern in testosterone concentrations is more clear?
Please define the meaning of the cross with the asterisk in the figure caption.

DISCUSSION
Line 325-326: Caution should be taken with this interpretation of the results. The authors have shown that testosterone, progesterone, cortisol, and androstenedione can be measured in dentin; however, only testosterone was used to assess whether hormones in dentin were biologically

relevant and could be reliably used in endocrine studies. The biological relevance of progesterone, cortisol, and androstenedione were not discussed and cannot be confirmed until additional studies are conducted.

METHODS

Line 447: Since each tusk was sampled differently, the methods section might be easier to read if the sampling process was divided by individual.

Lines 477-478: Please specify where the procedures were documented. Where is “elsewhere”?

Line 479: Did the samples that were used for hormone analysis have kerosene applied to them? Could that have an impact on hormone measurements?

Line 489-490: The authors note that most samples were 30 mg; however the tables provided in the Extended Data section note a wider range of measurements (eg. Line 827). Please clarify how much sample was used for hormone analysis.

Lines 511-514: Based on the tables provided in the Extended Data section (eg. Lines 827), it appears the sample mass used for each sample varied. Did the authors adjust the amount of extraction solution to keep the ratio of sample to solvent consistent? Ajó et al. (2022) found that sample mass significantly impacted hormone concentrations and recommended that sample mass to solvent ratios should be kept consistent. If the sample measurements did vary, I suggest conducting an additional study to determine if sample mass had an impact on the hormone results.

Ajó, A.F., Hunt, K.E., Dillon, D., Uhart, M., Sironi, M., Rowntree, V. and Buck, C.L., 2022. Optimizing hormone extraction protocols for whale baleen: tackling questions of solvent: sample ratio and variation. General and Comparative Endocrinology, 315, p.113828.

Lines 525-528: Were the samples run in duplicate? If not, how was the precision of the measurements determined?

EXTENDED DATA

Line 734: (Figure 6) How were the authors able to link the growth layers in each slab? Did the authors have possession of the entire tusk or just the slabs?

SUPPLEMENTARY DISCUSSION

Lines 842-843: Please include reference.

Lines 878-883: The pattern observed in testosterone data could also be due to differences in sample mass, if the mass of the samples or the sample to solvent ratio were inconsistent.

Lines 891-893: The authors are only able to conclude that testosterone measured in dentin is reflective of the animal’s hormone concentrations, since the results followed a pattern expected if an animal was experiencing an annual musth. The other three hormones have not been validated for their biological relevance.

Author Rebuttals to Initial Comments:

Reviewer #1:

This is a really excellent paper that would seem genuinely to open a new window on the biomolecular analysis of fossil or sub-fossil animals. As far as I am aware the analysis of steroid hormones in such materials using robust chromatography-mass spectrometry approaches is highly novel. I especially liked the strenuous lengths the authors have gone to test their approaches and ideas.

The progression from a modern elephant to fossil specimen is exactly the way to develop this type of work, rather than jumping straight into fossil specimens, as is commonly done. I also liked the analysis of the female elephant to lend further credence to the male reproductive patterns that had been revealed in the elephant and mammoth specimens.

Some caution is required in proposing the use of steroids as a sexing tool as this may be confounded by diagenesis. Perhaps this requires more careful testing or at least a caveat added as indicated on the annotated manuscript.

The amount of thought and effort that has gone into this paper is enormous and exquisitely documented in the extremely well-written manuscript. The illustrations do an amazing job at giving the paper life.

I know I'm not supposed annotate the manuscript but I have and have returned this to the editor. I am happy if this is shared with the authors. *(Reviewer comments transcribed from annotated pdf below – all Reviewer #1 comments are marked in blue)*. The comments I have made are entirely positive and only intended congratulate the authors on their exceptional work or raise helpful suggestions for improving the paper. There are only one or two minor technical corrections that are required. The authors have been unusually upfront about the analytical challenges faced in this sort of work and identified areas for improvement. Their data appear internal consistent and robust. A minor query I have is the structures of low mass ions detected in the MRM experiments following CID? Could this detail be added - I failed to find it in the literature I explored?

I have not commented on the endocrinological or ecological aspects of the paper has this is not my field. My comments focus on the fossil and analytical aspects.

This is the best paper in the field biomolecular paleontology or archaeology I have refereed for Nature or any other journal. I recommend publication of this manuscript.

We thank Reviewer #1 for their careful reading of our manuscript and their detailed comments. We also appreciate their validating feedback and praise. We are encouraged to know that other researchers are as excited and enthusiastic about these results as we are! Because the reviewer submitted comments in an annotated pdf, we transcribed them below and have responded to each comment in turn (in red).

Detailed comments below (copied/pasted from annotated pdf):

Line 102 “large amounts” – better to say “high concentrations”

We had been hesitant to use the word “concentration” *except when referring to the measurements from serum as is convention. However, to address the reviewer’s comment, we added a short sentence (a few lines before this) clarifying our use of* “concentration” *when referring to dentin measurements. The added sentence reads:*

New lines 96-97: “For dentin-derived steroid measurements, we use “concentration” to mean picograms of analyte per milligram of dentin.”

New lines 99-100: “large amounts” *changed to* “high concentrations”.

Lines 112-114 – Need to be cautious here as a poorly preserved tusk may display zero or low steroid cncentrations and be mistinterpreted as a female.

The reviewer’s concern here is warranted. Following Reviewer #2’s recommendation, this sentence has been removed.

Lines 122-124 – But this would have to beacked up by other indicators of excellent preservation - such as protein quality in the tusks.

This is the same concern that is stated above (lines 112-114) and we think it is adequately addressed by our revision there.

Line 234 – I don't see how this can effect the quantification - the specificty of the method used should get around interferences. I suspect degradation is the most like explanation in line with my comments above.

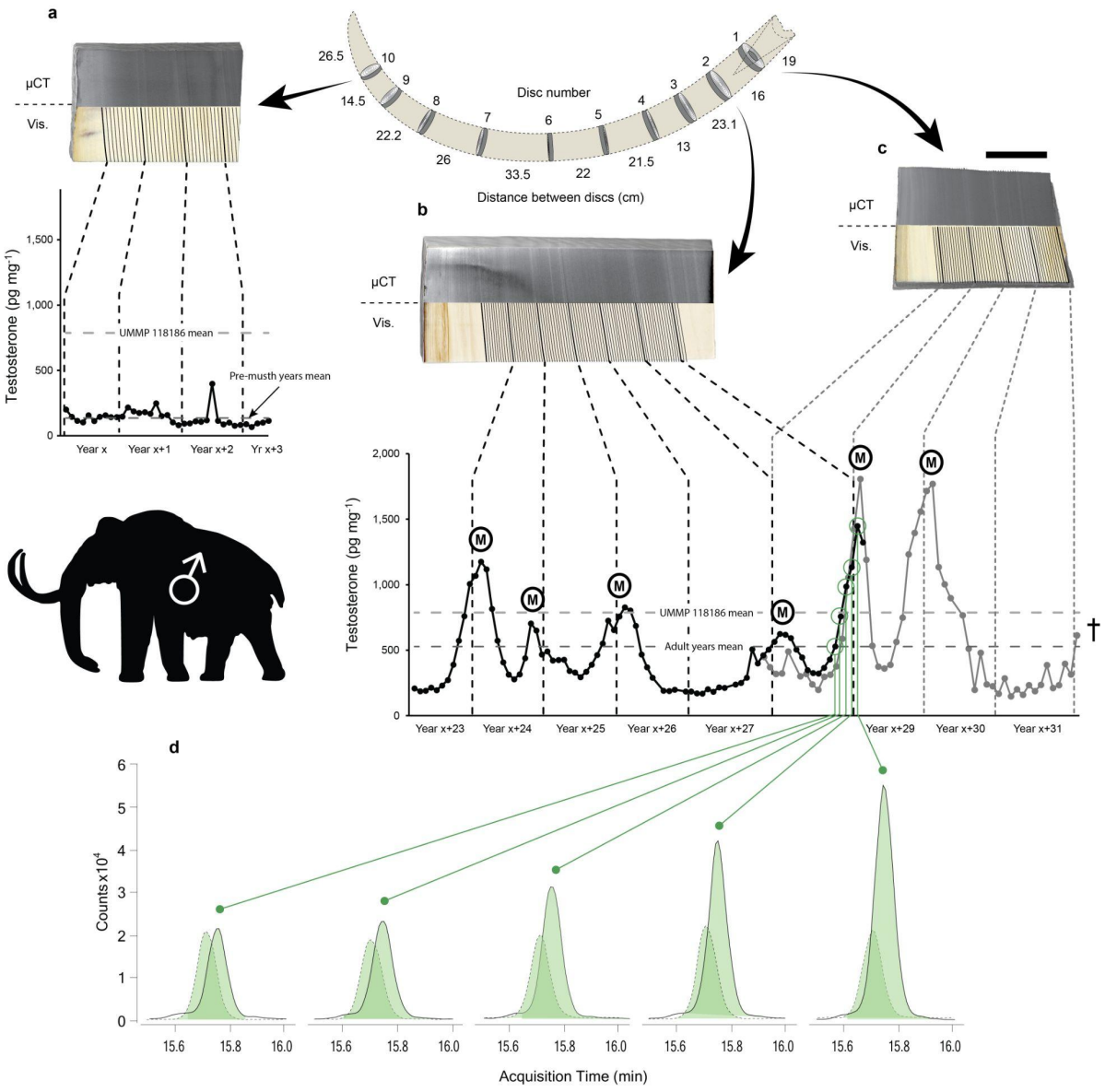
Here the reviewer recognizes the potential for lower testosterone measurements due to degradation (i.e., testosterone in the dentin actually is lower), but questions the relevance of analytical interference (i.e., testosterone just looks lower) in LC-MS/MS. Ion suppression due to other compounds that were not effectively removed during the Supported Liquid Extraction can interfere with measurement of peaks (raising the background ‘noise’ level in the region of the target ion peak). In other samples that we tested that had high organic content (e.g., pond-preserved remains), we saw low measurements that increased in the same samples when efforts were made to address the interference. We have revised the text for clarity:

New lines 230-233: “These lower values could be due to degradation of testosterone in the fossil tusk, ion suppression from organic compounds incorporated into the tusk dentin during diagenesis that co-elute with testosterone, or biological differences at the individual or species level.”

Line 250 – It would be nice to see the HPLC-MS/MS plots shown in f in Figure 1 displayed in this figure for the mammoth. I actually think given the novelty it would be appropriate to show the plots for the other steroids analysed.

We think this is a good suggestion and have modified Figure 2 to include HPLC-MS/MS plots to parallel the presentation in Figure 1. We have also added text to the caption for the newly added 'd' panel. The text mirrors the Figure 1 text and reads:

New lines 262-265: "d, LC-MS/MS plots of integration peaks for testosterone (solid line) and the deuterated testosterone internal standard that was added to each sample (dashed line) for five sequential samples that show the increase in the penultimate musth episode (See Extended Data Figures 1, 2)."



Line 286 – I really like the analysis of the female tusk as "control" for interpretations of the data from the male tusks.

Here and elsewhere we appreciate the validating comments from this reviewer!

Lines 318-319 – This is a very elegant dimension of the paper.

(See above)

Lines 325-326 – I agree - this very exciting.

(See above)

Line 501 – a mass is not an aliquot it is a proportion.

We were using "aliquot" simply to mean a portion of a larger whole and think that the use is appropriate here. From www.oxfordlearnersdictionaries.com: "a small amount of something that is taken from a larger amount, especially when it is taken in order to do chemical tests on it". Looking further into this definition, we have found examples where "aliquot" is said to refer to a "known fraction" of an original mass, but in none of these cases is there a rationale for why the particular fraction involved matters. We are left with the impression that some precedent can be cited in favor of both approaches, but most instances we have seen resemble our original use of "aliquot".

5-mg aliquots can therefore be removed from larger powder samples.

There are other words that would work here (e.g., "subsample"), but we feel that "aliquot" is the most appropriate without additional revisions, and so we have left this without revision.

New line 529: *we changed* "samples" *to* "These aliquots" *to clarify reference to the 5-mg subsamples as the portions that were pretreated for isotope analysis.*
New line 532: *wording revised to eliminate the duplicate use of* "aliquot" *for the 1-mg portions transferred to the mass spectrometer for analysis. New text reads:* "... 1-mg portions of each pre-treated powder sample…"

Lines 510-554 – too many uses of the word sample in this section - call them what they are to avoid confusing the reader.

We have revised to eliminate a few repetitions of the word "sample" for readability (including accepting the suggestion for line 521 below). We also made a few other minor wording adjustments to this paragraph for readability.

New lines 538-540:
"Pipetted" *replaces* "added"
"An extraction solution" *removed*
"A 40 % methanol solution" *replaced with* "40 % methanol"
Sentence divided to separate pipetting methanol and internal standard (ending with "into the 2-ml centrifuge tubes containing powder samples"):
"To extract steroid analytes from powder samples, we pipetted 1.5 ml 100 % HPLC-grade methanol and 100 µl internal standard made from deuterated forms of each analyte in 40 % methanol into the 2-ml centrifuge tubes containing powder samples."

… from addition of beads (beginning subsequent sentence with "We then added…" and removing "2-ml centrifuge" from that sentence) which we combined with the subsequent sentence:

"We then added ~250 mg Lysing Matrix D beads (MP Biomedicals) to each tube to prevent clumping during agitation and rotation."

Line 520 "tert" – should be italicised

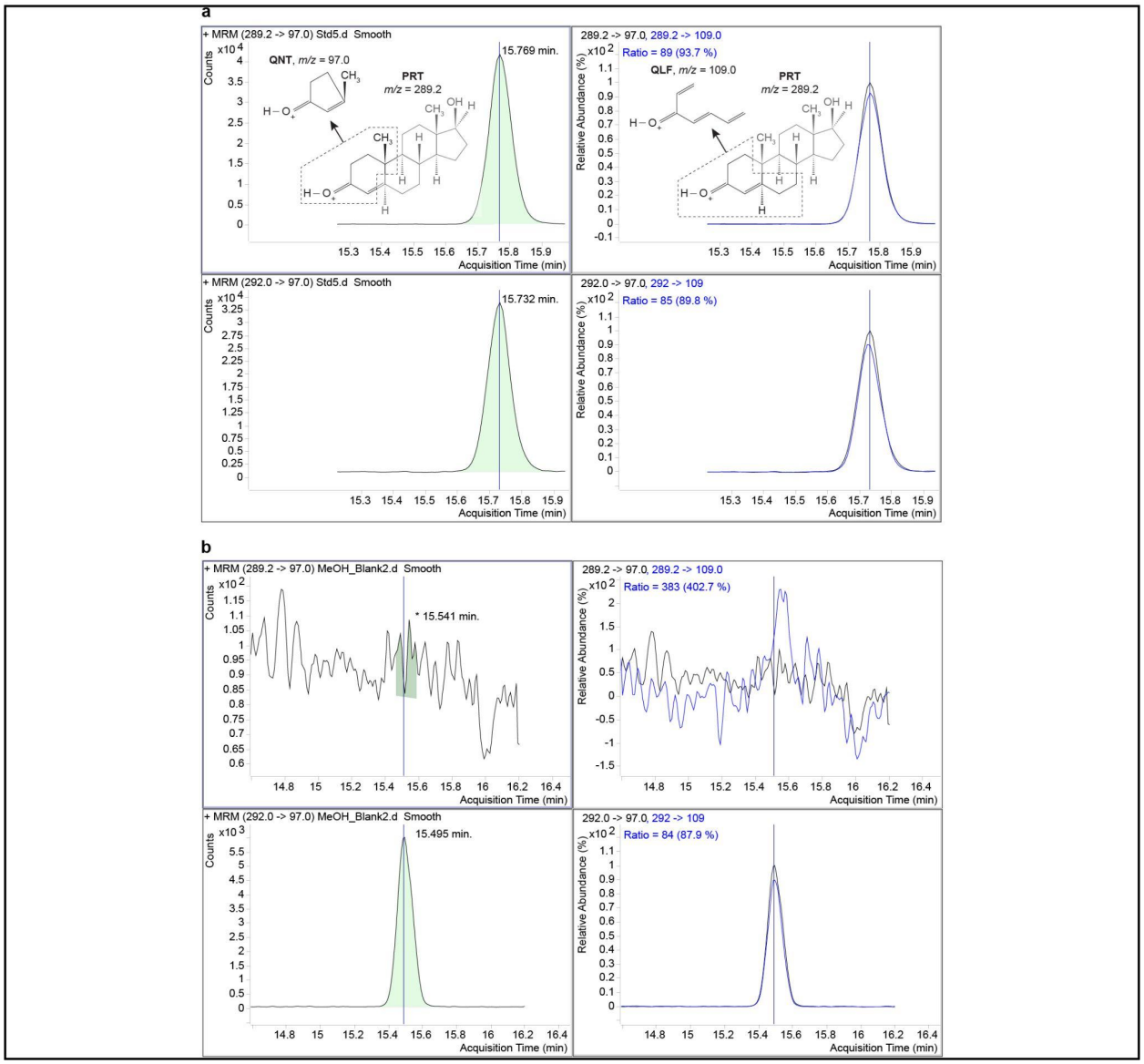
New line 547: *We thank the reviewer for catching this error and have revised accordingly.*

Line 521 "samples" – replace with extracts

New line 548: *Reviewer suggestion accepted and revision made. We also removed the subsequent "sample" in the same line (that we deem to be unnecessary).*

Line 524 reference 47 – I looked at this paper but it does not contain full details, especially as what some of the ions studied correspond to in terms of structure.

Reference 47 (Wright et al., 2020) documents the delta-4 steroid analytical method used in our lab and includes necessary parameters for duplicating these analyses. Equipment, analytical columns, calibrators and internal standards, chromatographic scheme, steroid-specific MRMs and qualifier ions, response times, LLOQs, solvents, flow rates, and other LC-MS/MS settings are documented therein. The chemical structures of MRMs for the steroids are not documented in that reference and are typically not referenced in LC-MS/MS reports. In some cases, such as for testosterone (289 m/z with quantifier ion 97 m/z and qualifier ion 109 m/z), the structure is now considered established knowledge and typically not documented or referenced, but to address the reviewer's comment, we have added structure diagrams for this fragmentation to Extended Data Fig. 1a.



Line 531 “destruction” – I think the correct term is dissociation

New lines 558-559: *We agree with the reviewer that* “destruction” *is not the best word here. However, rather than using* “dissociation,” *we prefer the more technically accurate term* “fragmentation,” *and we have revised accordingly.*

We have also added a reference here to the revised Extended Data Fig. 1a to

Line 532 “integration of” – integration of the peak area

New line 559-560: We accept the reviewer’s suggestion here and have revised for clarity.

Lines 542-547 – Some reviewers might worry about this but achieving exhaustive extractive (100%) is difficult to achieve. Adopting a consistent extraction method for all tusk tissue samples is appropriate for making comparison between growth phases within a tusk.

As the reviewer recognizes, there is no need to ensure exhaustive extraction, and demonstration of complete extraction is a challenge we do not attempt to take on here. Use of a consistent extraction procedure produces comparable results. Below, we address a related comment from Reviewer #2 concerning the use of unequal sample masses, which for the range of masses we used should not be a problem.

Lines 549-550 – Don't they have this under control - I thought this was implied from the above - perhaps word differently.

We think that the reviewer's confusion is understandable here. Our wording was unclear because our argument above was intended to be limited to within-specimen consistency in order to provide interpretable patterns of change in an individual growth record. For comparing absolute abundances of various steroids in dentin to evaluate sex of different individuals, the comparison would more heavily rely on consistent extraction completeness, a factor that we think could be impacted by different preservation. However, we recognize that this would require further explanation and is beyond the scope of what we think is important to convey. This sentence has been removed and minor additions were made to the previous sentence to accommodate the change:

New lines 574-575: "..., these experiments were designed only to detect temporal changes in steroids preserved within each tusk record."

Lines 550-552 – I'm not sure I'm following these caveats.

To address the reviewer's uncertainty, we have added clarification that the normalisation of records enhances between-specimen comparisons (within-specimen changes could be meaningful even with significant degradation and without normalisation), but that it does not appear to be necessary for the comparisons made in this study.

New lines 575-579: "Normalising testosterone records relative to series minimum (Extended Data Fig. 9, Supplementary Discussion) and relative to androstenedione measurements (Extended Data Fig. 10, Supplementary Discussion) may produce more comparable datasets from specimens with differential preservation, but for this study, the results are clear without normalisation."

Lines 557-558 – I think additional references are needed in the methods section - this reference is quite light on the detail I would have expected.

Two other references are already cited in the methods section, but are already listed in the references for the main text and were not repeated here. We have added a citation for "sample planning," but it is the same reference cited for "thin sections" (Fisher et al., 2014) that is already

listed in the main reference section. Other than the structure of MRMs (addressed above and with a revision of Extended Data Fig. 1), we are not certain what additional details the reviewer is hoping to see.

As this is not primarily a methods paper, we think that citing references for the LC-MS/MS instrumentation/procedure, sample planning, thin sectioning, and stable isotope analysis should be adequate. Sample collection is simple and does not require the additional explanation that a reference would provide. Without citation, we detail CT scanning equipment/procedures and the sample treatment procedure that we developed for LC-MS/MS analysis of tusk dentin samples (this procedure was used for various hair extraction procedures as a starting point but is different enough that adding a hair analysis reference would unnecessarily complicate the description). If Nature *requires additional background references for the methods section, we will augment as necessary.*

Line 696 “levels” – Not sure I like the use of the word level here or anywhere else in this paper.

The reviewer does not clarify their objection to the use of “level,” but we recognize that it is a nontechnical word and that use of “concentration” and “abundance” could be substituted where appropriate. In the originally submitted manuscript, we used the word “level” 25 times in reference to hormone abundances/concentrations. We have made the following changes:

New line 156: a… concentration
New line 233: patterns
New line 242: *(deleted)*
New line 283: *(deleted), made verb singular*
New line 295: value
New lines 568-568: “Steroid concentrations for tusk dentin samples were calculated by adjusting measured concentrations according to sample mass to produce results expressed in pg mg-1 dentin powder.”
New lines 699, 725, & 746: records
New line 750: *(deleted)*
New lines 820, 826, & 844: concentrations
New line 915: concentrations
New line 986: *(deleted)*
New line 1023: values

Lines 818-820 – I like this idea.

We do too! And as stated above, we appreciate the validation.

Lines 871-893 – This is perfect.

(See above)

Lines 899-901 – I agree

(See above)

Lines 913-917 – this is odd - I admire their honesty.

As one of the first studies of this kind, we did not expect consistently usable data. Here, we document inconsistencies of unknown origin that provide obstacles to interpreting the patterns in progesterone and cortisol as being biologically meaningful. However, we regard the measurements, particularly for progesterone, as indication that these hormones are preserved in dentin and therefore want to document their presence in our samples. Further work is needed to identify the source of inconsistent results and to develop methods that will produce consistent results that provide bases for physiological interpretations. We have revised the text with additional details to improve clarity. We have added some speculation about the cause of inconsistencies in our revised Supplementary Discussion section titled Dentin steroids **(New lines 931–982)**.

Lines 993-997 – I have no idea as to the ethics of this? Ivory is a contraversial matter so some exploration as to how this works with international conventions and greements would be worth persuing and stating.

We appreciate the reviewer's perspective on the controversial nature of elephant ivory and have added the following text explaining the history of the tusk and the timing of some relevant treaties and laws regulating trade in wildlife parts, including elephant ivory. This text clarifies that the tusk was legally obtained and donated to the museum for research, which has removed it from any potential involvement in the current illegal ivory trade.

The new text reads:

New lines 1081-1094:

“**Specimen acquisition**

This study was undertaken with one of the goals being to increase our understanding of African elephant biology in support of conservation. The African elephant ivory analysed in this work (UMMP 118186) originates from an animal that was hunted as part of a safari near Maun, Botswana in September and October of 1963, and the tusks were imported to the US as trophies soon after the safari. At the time, treaties and laws that now protect elephants through restrictions on trade in their parts (including ivory), such as the 1973 Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), the Endangered Species Act of 1973, and the African Elephant Conservation Act of 1988, had not yet been developed. The ivory in question remained in the family of the hunter and was never sold after it was imported. It was donated by a descendent of the hunter to the University of Michigan Museum of Paleontology for use in scientific research, and the donor received no compensation from the

museum. The two mammoth tusks, UMMP 118711 and UMMP 118710, were collected and exported following relevant laws and with appropriate permits."

Reviewer #2:

The authors present the first evidence of musth in a woolly mammoth by examining testosterone patterns in ancient tusks. Specifically, the authors measured testosterone, progesterone, cortisol, and androstenedione in dentin from an adult male African elephant, an adult male woolly mammoth, and an adult female woolly mammoth. Testosterone patterns in male tusks followed a pattern that would be expected if an animal was experiencing an annual musth. Additionally, testosterone peaks were absent in the female tusk and in a portion of the tusk representing pre-musth years of the male mammoth, adding strength to the findings. The work presented here is the first successful biological validation of dentin hormones and is not only valuable for studying historical specimens but also for assessing endocrine patterns in tusks from modern species.

I applaud the authors for their work; however, I have comments that should be addressed before publication. In particular, the authors need to clarify the sample masses that were extracted for hormone analysis. 30 mg is mentioned in the paper; however, the sample masses listed in the Extended Data section vary greatly. Recent research has shown that sample mass can impact apparent hormone concentrations (Ajó et al. 2022). If the sample masses listed in the Extended Data section are correct, I suggest that the authors conduct an additional analysis to ensure that the mass of the samples does not impact their results. This could be done by weighing out the same sample (or a pooled sample) in range of masses, extracting them using the protocol provided in the paper, and determining whether sample mass impacts their results.

The progesterone and cortisol data included in this paper are interesting; however, the results are only briefly discussed in the Results and Extended Data section and I question whether they add value to the question of musth in woolly mammoths. I suggest that the authors either include additional information on how these hormones are relevant to musth or potentially remove them from the study, if no relevance can be determined.

Additionally, caution should be taken when interpreting the main findings of this study. Although the authors have shown that testosterone measurements in dentin are biologically relevant, as testosterone was shown to peak annually, which is indicative of an annual musth; biological validations for progesterone, cortisol, and androstenedione were not included. Therefore, the authors cannot conclude that they are biologically relevant.

Tremendous job and I am looking forward to the final version of this manuscript.

We greatly appreciate Reviewer #2's thorough review and comments! After working our way through all of the requested revisions, we think that we have adequately addressed all concerns. We are confident that these revisions have enhanced the clarity of the manuscript. Specific responses to all detailed comments are included below (marked in red).

Detailed comments below *(all Reviewer #2 comments are marked in black)*:

ABSTRACT

Line 27: More appropriate references for this sentence:
McCormick, S.D. and Romero, L.M., 2017. Conservation endocrinology. BioScience, 67(5), pp.429-442.
Wikelski, M. and Cooke, S.J., 2006. Conservation physiology. Trends in Ecology & Evolution, 21(1), pp.38-46.

We thank the reviewer for these suggestions. We will cite the first reference (McCormick et al., 2017) in **New Line 26** *as a review article that helps establish broad significance for our study and provides background. We are keeping a reference for circadian rhythmicity in hormone concentrations as one that provides an example of immediate steroid concentrations from serum as a contrast to the time-averaged concentrations that can be measured from appositionally formed tissues.*

Line 28: It is still debated whether steroids in hair reflect historic endocrine patterns (see paper below). I suggest using baleen as an example of a matrix that can provide historic endocrine data.
Kalliokoski, O., Jellestad, F.K. and Murison, R., 2019. A systematic review of studies utilizing hair glucocorticoids as a measure of stress suggests the marker is more appropriate for quantifying short-term stressors. Scientific reports, 9(1), pp.1-14.

Accepting that there are questions surrounding the relevance of hair-steroid studies, we do not wish to weigh in on this point here. Instead, our intent is to acknowledge other work, much of which concerns hair studies. Thus we are keeping the hair reference in **New line 28***, but have added a citation to another already referenced article (Hunt et al., 2014) and changed the wording from* "hair" *to* "keratinous tissues."

INTRODUCTION

Line 53: Include the scientific name of Asian elephants.

We appreciate the reviewer catching this omission and have added the scientific name to the text in **New line 52**

Lines 57-59: Provide reference.

Citation added to **New line 58** *for already referenced article (Fisher, 2018)*

Lines 67-71: I suggest including the rationale for analyzing the other hormones presented in this paper here.

The steroids reported here are the 4 repeatedly detected in a 24-steroid panel regularly run on serum samples in our lab. We did not set out to analyze these specific steroids, but seeing a strong signal of testosterone we decided to test the use of tusk records for identifying musth. Despite this opportunistic focus on testosterone, we consider the presence of other steroids

worthy of mention. To clarify this, we made a few minor revisions to the final paragraph of the introduction.

New lines 70-72: "We used fine-scale serial sampling of dentin paired with liquid chromatography tandem mass spectrometry (LC-MS/MS) to survey steroid hormones in tusks. Using the LC-MS/MS results, we identified the occurrence and timing of intervals of increased testosterone …"

RESULTS

Lines 86-87: The scientific name for African elephants and woolly mammoths were provided in the introduction; you can remove them from this section.

We have removed scientific names from **New lines 73, 82, 83, 129, 191, & 270.** *Scientific names are now only listed once each in the main text.*

Lines 107-110: Are the ranges for the testosterone and androstenedione ratios in dentin provided in the supplementary text?

Ranges for the testosterone:androstenedione ratio in serum (0.4 – 9.8) and dentin (0.4 – 10.7) are listed in **New lines 105-106**. *All steroid measurements are contained in the supplementary information.*

Lines 104-106: Need references for the effects of oxidation and bacterial degradation of tusks.

New lines 101, 102: *We have added the following references:* Ridlon et al., 2013 *for the bacterial degradation of cortisol.* Davio et al., 2020 *for the oxidation of cortisol.*

Lines 113-114: I suggest removing this sentence as the rest of the paragraph discusses how sexing an individual based on their hormone measurements would be difficult.

Reviewer #1 also had reservations about this sentence, and we agree that it did not add significantly to the manuscript, so we removed it.

Line 125: (Table 1) The female African elephant is noted to have a median testosterone concentration of 0 in Table 1. Can the authors elaborate as to why they think they got this result? Was the concentration simply lower than the LLOQ?

New lines 121-125: *This is a mistake. The reviewer is correct and it should read "NQ." We have revised the table accordingly and added the abbreviation NQ (not quantitatable) to the caption.*

Line 133: The scientific name for African elephants has already been mentioned – please remove.

Scientific name has been removed here.

Line 137: I suggest adding a brief sentence describing the rationale for analysing oxygen isotopes.

We appreciate the reviewer's suggestion and revised the first sentence of the paragraph to explain the reason for investigating stable isotope records and address the carbon isotopes (values are included in the supplementary information) and explain why they are not discussed. The revised text reads:

New lines 132-134: "To help evaluate the season of musth episodes in the tusk growth record, we analysed stable isotope values from dentin carbonate. The serial record of carbon isotopes (^{13}C) lacks clear seasonality, but oxygen isotope ratios (^{18}O) show a roughly sinusoidal pattern"

Lines 137-139: Reference "Fig. 1" here.

New line 136: *We agree with the reviewer and have added a reference to* "**(Fig. 1)**" *at the end of the sentence.*

Line 145: I do not see a CT scan of the complete specimen in the supplementary section. Only the CT scans of the 10 individual sections were provided. Were the authors able to count all the growth layers in the tusk using the CT data and confirm that the core taken from the tusk truly represents the last 9 years of the animal's life?

The CT scan data for the complete tusk of the African elephant (UMMP 118186) was not actually included with the initial submission. ROIs for scans in the University of Michigan library's digital data repository are included in this revised draft. The "10 individual sections" referenced by the reviewer here were not in fact CT data, but instead were thin sections from the ten discs that make up the only parts of the woolly mammoth specimen (UMMP 118711) that were accessioned in our collection. The core in question contains nine growth increments visible under reflected light and using microCT imaging. The oxygen isotope data strongly support these being annual increments, and the core extends to the pulp cavity surface. Thus, the core contains the final nine years of growth in the tusk, which stopped growing at the time of death. The full tusk CT scan shows that these nine years of growth account for roughly 30% of the tusk's length.

Line 151: I'm curious as to why the authors did not core a section of the adult male elephant tusk at the distal end (presumably pre-musth years) to replicate the work on the male woolly mammoth tusk?

In the case of the woolly mammoth, the distal samples were useful for confirming that the testosterone pattern in the proximal years was biologically meaningful (the musth perturbations only showed up where expected, after the animal became a fully mature adult). However, in the

case of the elephant, the 9 years of tusk growth already include 5 pre-musth years (based on the lack of characteristic testosterone fluctuations–this is the same as what we would expect to see in any more distal section of the tusk record) followed by the first 4 years of annually recurring musth. The pattern in the elephant tusk record matches expectations based on previous studies that document testosterone changes during musth in modern African elephants. Thus, we are confident that the modern elephant testosterone record includes direct evidence of musth episodes, and that we have sufficient data from earlier years that more distal samples would not alter interpretations.

Line 157-159: Link this sentence to the oxygen analysis results.

Here we accept the reviewer's suggestion. Furthermore, we use the revision as an opportunity to eliminate some redundancy in the original text. The association with the Okavango 'flood' season with the broader region's 'dry' season was documented above (lines 140-142). In this more streamlined revised text, we simplify the time of year for musth episodes ("regional dry season") and link to the oxygen results:

New lines 155-156: "... during the regional dry season when tusk ^{18}O has rebounded from rainy-season lows" *replaces* "approximately coinciding with the 'flood' season in the Okavango Delta (which overlaps with the 'dry' season for the broader region)"

Line 194: Scientific name already provided.

Scientific name has been removed here.

Line 258: Please elaborate on the features that were used to correlate growth layers between each slab.

The reviewer's request is reasonable, and we made a few minor revisions here that we think should be adequate for this figure caption. First, we added text ("of growth features" and "by visual evaluation of") and referenced Extended Data Fig. 6, which shows the thin-section correlations. The new text reads:

New lines 254-256: "Overlap of plots b and c in year x+28 is based on correlations of growth features identified by visual evaluation in thin sections of discs 1 and 2 (Extended Data Fig. 6)."

We also revised the text of the Extended Data Fig. 6 caption to elaborate on the approach used to correlate the features in thin sections of adjacent discs. The following sentence was added to the Extended Data Fig. 6 caption:

New lines 768-772: "Corresponding features in thin sections of adjacent discs were identified from matching qualities such as thickness, spacing, and fine-scale intra-feature patterning of light and dark bands (many annual 'features' are made up of multiple growth lines)."

Line 265: (Figure 2) Please define the cross symbol in the figure captions for all figures.

In the captions for each figure, we have added the following text:

New lines 178-179, 261-262, 304: "The dagger symbol marks the end of life."

Line 266: The heading title for the female woolly mammoth is missing. The male elephant has a heading title on line 131 and the male woolly mammoth has a heading title on line 189.

New line 268: *The heading* "**Female woolly mammoth**" *must have been inadvertently deleted during final editing prior to original submission. We have added it back in.*

Lines 267-268: Add "female" to this sentence.

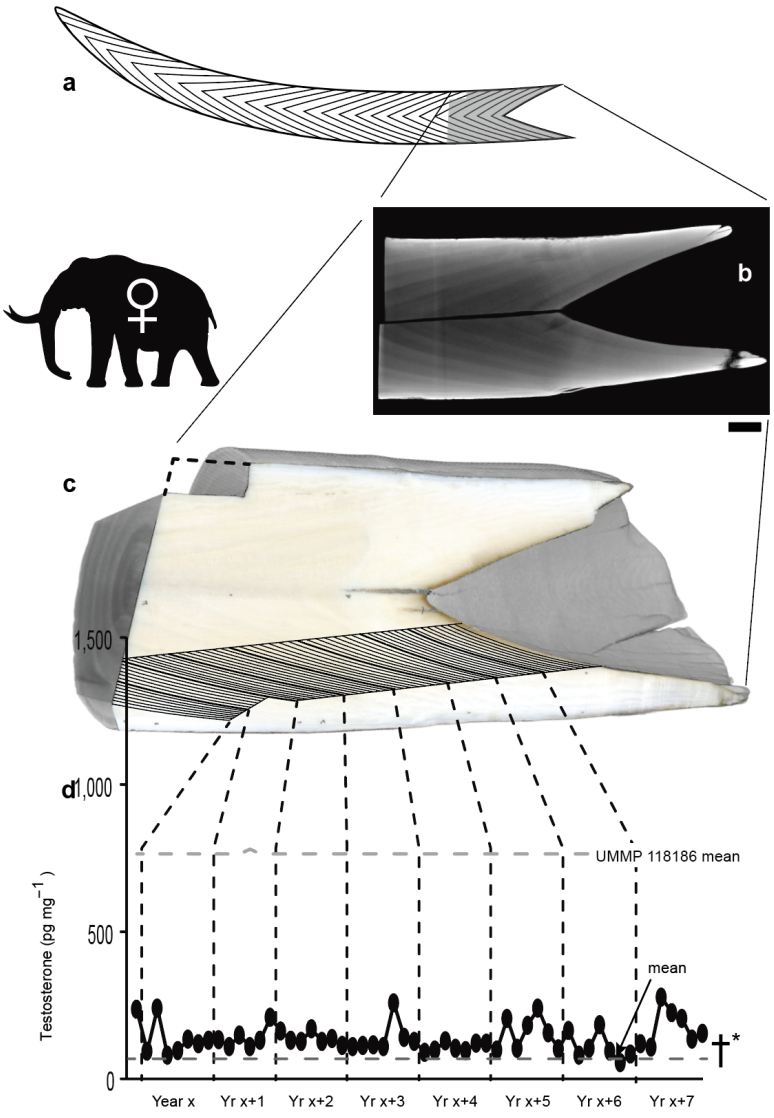
With the heading now included, we think the reviewer might find this less necessary. Furthermore, that it is in fact a female tusk is an interpretation based on diameter and is addressed two sentences later on **New Line 272.** *Thus, we declined to make this change.*

Line 277: Change "was" to "were".

Here the singular subject ("A series") on **New line 279** *requires a singular verb ("was" rather than "were"), so we have left it unchanged.*

Line 284: (Figure 3) Could the authors adjust the figure so that the pattern in testosterone concentrations is more clear?

We regard the consistent low values to be the most relevant information here. The chart y-axis was intentionally scaled similarly to those showing musth patterns in male tusks and displays the same range (0-1,500) as the axis for the chart of the juvenile male mammoth years. With this scale, the plot clearly shows the female testosterone record as being consistently much lower than in the male records (and comparable to the juvenile male record). At the very least, we think that it is important to include the UMMP 118186 mean in the plot, which requires an axis that includes 0-~800. The y-axis could be scaled with the UMMP 118186 mean (788) as the maximum value, but even then, unless the height of the chart is increased dramatically, there is little to no benefit for showing the 'pattern' of the values (see below).



To address the reviewer's concern, we added two sentences in the figure caption to explain our choice for the axis scale. The added sentences read:

New lines 296-300: "The vertical axis is scaled to include the male elephant (UMMP 118186) mean (787.6 pg mg^{-1}) and to facilitate comparison to the charts in Figs. 1 and 2. This scaling highlights the consistently low values in the female mammoth testosterone record (relative to adult male values) while demonstrating the absence of the musth pattern present in the adult records for the African elephant (UMMP 118186) and male woolly mammoth (UMMP 118711)."

Please define the meaning of the cross with the asterisk in the figure caption.

This has been done. Addressed above.

DISCUSSION

Line 325-326: Caution should be taken with this interpretation of the results. The authors have shown that testosterone, progesterone, cortisol, and androstenedione can be measured in dentin; however, only testosterone was used to assess whether hormones in dentin were biologically relevant and could be reliably used in endocrine studies. The biological relevance of progesterone, cortisol, and androstenedione were not discussed and cannot be confirmed until additional studies are conducted.

We appreciate the reviewer's caution and recognize the boldness of our assertion here. While we think our assertions are justified, we see value in a slightly more qualified statement. The reviewer is correct to point out that not all of the four steroids detected were shown to be biologically relevant. However, it was not only testosterone, but also androstenedione that reflected musth episodes (this is now better explained in the Dentin steroids section of the Supplementary Discussion). To revise the present statement, we omitted the word "reliable." Below in the same paragraph, we state "With reliable results for some steroids," which is a more careful and specific statement, so "reliable" is not necessary here. We also omitted the word "dramatic" feeling that it is unnecessary and could be what caused the reviewer to react with caution. We have left the rest of the sentence intact and think our optimism is justified. It now reads:

New line 334-335: "This study establishes dentin as a repository for modern endocrine studies and sets the stage for advances in the developing field of palaeoendocrinology."

METHODS

Line 447: Since each tusk was sampled differently, the methods section might be easier to read if the sampling process was divided by individual.

Here the reviewer appears to be referring to the CT scanning of each specimen rather than the sampling. In fact, the only section of the methods that is specimen dependent is the MicroCT. As such, we don't think that separating the methods for each specimen would be beneficial. We have left this unchanged with Sample planning, Thin sections, Sample collection, Stable isotope analyses, and Steroid analyses all detailing methods applicable to all specimens/individuals. The MicroCT section has details listed in sequence for the UMMP 118186 core, the complete tusk of UMMP 118186, sample slabs from UMMP 118711, and the complete specimen of UMMP 118710.

Lines 477-478: Please specify where the procedures were documented. Where is "elsewhere"?

The reference is already provided but must have been overlooked by the reviewer. These procedures were documented in Fisher et al, 2014, and the citation immediately follows the sentence **(New line 502)** *"... following procedures documented elsewhere."*

Line 479: Did the samples that were used for hormone analysis have kerosene applied to them? Could that have an impact on hormone measurements?

Kerosene was only added to thin sections being examined using optical microscopy. These thin slices of dentin were separate from the sample slabs. This should already be clear from the section on ***Sample collection*** *that states:*

New lines 518-522: "To limit the potential for contamination we cut specimens without lubrication, we handled all prepared slabs with gloves, we kept sampling surfaces clean and dry throughout the process, we wiped all tools with HPLC-grade methanol between each sampling, and we used a clean scalpel to scoop and transfer powders directly from the mill into 2-ml centrifuge tubes used for methanol extraction."

We declined to add clarification to the ***Thin sections*** *section because it seemed out-of-place for discussion of sample slab treatment.*

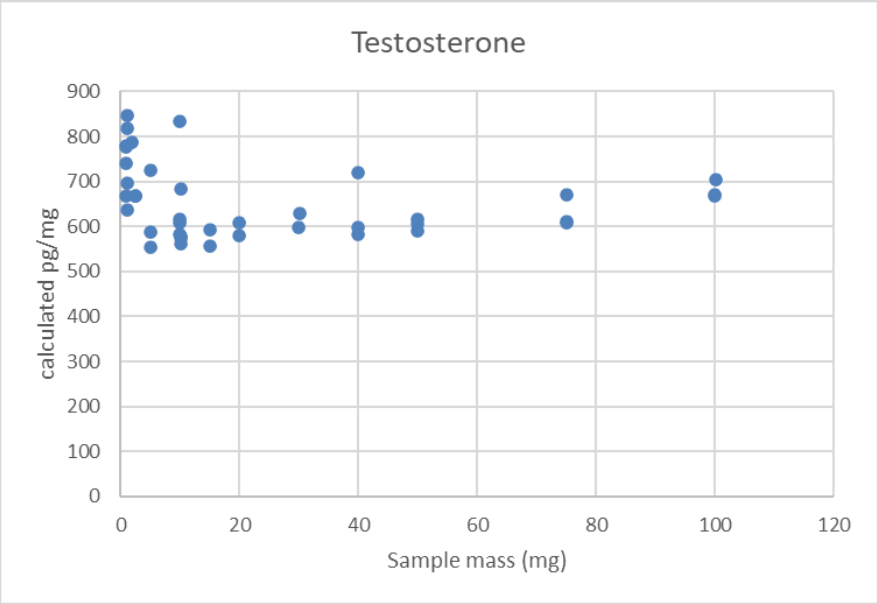
Line 489-490: The authors note that most samples were 30 mg; however the tables provided in the Extended Data section note a wider range of measurements (eg. Line 827). Please clarify how much sample was used for hormone analysis.

The reviewer makes a good observation here. The methods section documents the sample mass that we settled on for standard practice after many measurements over the course of the study. However, many of the earlier-collected samples were significantly larger than 30 mg. We did not measure out a consistent mass for each sample. Instead, precise masses were measured on a microbalance, and all LC-MS/MS results were corrected for that sample mass. Our decision to use the entire mass of sample collected was to avoid subsampling issues (in early tests we were having difficulty guaranteeing powder homogeneity) and to limit the number of vials/tubes/tools used to limit the potential for contamination. However, we were only able to roughly estimate the mass of each planned sample, so we settled on targeting "about 30 mg."

The reviewer does not specify the source of the published "mass effect" here but later mentions Ajo et al 2022. Studies of fecal and keratin-bound steroids over the last 20 years have seen a mass effect where smaller samples have apparent higher concentrations than larger samples. The reviewer questions whether the "musth pattern" could simply be a mass effect. All of the studies that noticed a significant impact of sample mass (under about 10 mg) used immunoassays for their measurements and one 2010 study (Hayward et al.) was able to eliminate the effect by 1) using a precise scale and 2) completely evaporating the ethanol before reconstituting with the assay buffer (rather than diluting the methanol with the buffer). Hayward et al. concluded that the ethanol that was usually left in the sample and mixed with the buffer was interfering with the assay. The most recent contribution to this discussion was the 2022 paper by Ajo et al. that concluded with the recommendation of using sample sizes of at least 20 mg (they saw the start of a mass effect in 10 mg samples, and 20mg was where it disappeared... they didn't test anything between 10 and 20 mg).

Our samples were all above 10 mg (only 2 were less than 20 mg) and we did not use immunoassays, so it's not clear that we should be concerned about a mass effect in our study. Furthermore, within-slab calculations show no significant correlations between sample mass and testosterone -- all p-values were more than 0.05, and the highest R-squared value was 0.027.

However, taking the reviewer's concern at face value, we decided to do a controlled test to look for a mass effect using our procedure. We milled a large bulk sample from the tusk slab for UMMP 118186 (African elephant), tumbled the powder overnight to homogenize it, and ran a series of tests using subsamples at various precisely measured masses (target mass +/- 0.1 mg) to see if results were correlated to sample mass.



(NOTE–We are currently at the maximum 10 Extended Data Figures. However, with editor approval, this figure could be formatted as an 11th Extended Data Figure.)

In these tests, we did see a limited "mass effect." It appeared only in the lowest mass samples (most visible in 1-2.5 mg samples) and in the most extreme cases (1-mg samples produce an average testosterone measurement about 20% higher than samples > 10 mg) could not account for the magnitude of variations in the male testosterone records. The source of this mass effect is unclear, but notably, it contrasts with the 4-5x increases at low masses seen from immunoassays reported in the literature. There is also a little more variability in results from samples less than 15 mg, which is expected, both due to sampling from an imperfectly homogenized bulk sample and increased impacts of measurement and curve-integration errors in LC-MS/MS. It is also worth noting that the 1-mg sample concentrations (6-9 pg/ml) are below the 10 pg/ml LLOQ for testosterone based on previous tests. In addition, values for the 1 to 2.5-mg samples are below that of the lowest standard used to create the calibration curve (~24 pg/ml). So, the measured effect may just result from the method not providing reliable results for such low concentrations. Fortunately, the lowest pg/ml measurements for the adult male records

were in the 40-60 pg/ml range and the vast majority were in the hundreds (pg/ml) range. Most results from the juvenile male and adult female mammoth samples were below 100 pg/ml, but even in those cases, almost all were above the lowest calibrator value.

To further investigate this issue, we also ran several 1-mg samples using differing amounts of solvent (to see if steroid extraction from larger samples was less complete than extraction from smaller samples due to solvent saturation effects). However, we saw no difference between 1-mg samples using 200 ul MeOH and those using 1600 ul MeOH.

Finally, although we do not detail the specifics here, the patterns are similar for androstenedione. The results from progesterone were less consistent overall and showed no clear mass effect. As in most other samples from this tusk, this bulk powder sample doesn't show cortisol, so we were not able to evaluate whether sample mass affected cortisol measurements similarly.

Because this is not a methods paper, because we are currently at our limit for Extended Data figures, and because further tests would be needed to do justice to the issue, we hope that you are satisfied with us making the following revisions:

1) *In the methods section titled Sample collection, we replace "Most samples were about 30 mg, a mass that consistently produced clean signals." with more detailed and accurate documentation:*

New lines 513-517: "Sample masses ranged from 12.8 – 86.7 mg with only two less than 20 mg and the majority between 30 and 50 mg. Over the course of the study, we settled on a target mass of approximately 30 mg, a mass that satisfied practical considerations for time and effort while producing consistent results. We saw no evidence that sample masses impacted our results (see Supplementary Notes for more detail)."

2) *The following section has been added to the Supplementary Notes:*

New lines 1031-1054:
"**Sample mass**
We are aware of an apparent sample mass effect in other studies using immunoassays to measure hormones in fecal samples and baleen.(Millspaugh and Washburn, 2004; Hayward et al., 2010; Ajo et al., 2022) We do not expect a similar *mass effect* in our analyses that used LC-MS/MS rather than immunoassays. Nevertheless, in published results, samples above 10 mg produced consistent results independent of sample mass, and all of our samples were above 10 mg (all but two were greater than 20 mg). In our tests, samples as small as 1 mg routinely produced clean signals of testosterone, androstenedione, and progesterone (cortisol was generally inconsistent), but concentrations in samples less than 5 mg were often below the pg ml^{-1} concentration of the lowest calibration standard and therefore could not be confidently quantitated. Some of the smallest samples provided signals for testosterone that were clean but were below the LLOQs previously established in controlled tests in the lab using the same

equipment and procedure (Wright et al., 2020), and were thus considered unreliable. Additional work would be necessary to rule out a mass effect at the lowest sample masses, but as reported by other authors, we found no mass dependency of results in samples larger than 10 mg. Furthermore, if a mass effect were cause for concern in this study, we would expect to see a high correlation between sample masses and testosterone measurements (i.e., a correlation coefficient close to +/- 1.0 with $p < 0.05$). Because sample masses tended to be fairly consistent within each sample slab, and because we were interested to see if sample masses for each slab could have influenced its steroid series patterns, we calculated an independent Pearson correlation coefficient for each slab. The best of the within-slab correlation coefficients was -0.163 ($p = 0.23$, $n = 56$) and the within-slab correlation with the lowest p-value had a coefficient of 0.155 ($p = 0.17$, $n = 79$). With correlation coefficients close to 0.0 that are not unlikely to be random (they have high p-values), we see no evidence that a sample-mass effect on steroid values produced the testosterone patterns we use to interpret musth episodes."

Lines 511-514: Based on the tables provided in the Extended Data section (eg. Lines 827), it appears the sample mass used for each sample varied. Did the authors adjust the amount of extraction solution to keep the ratio of sample to solvent consistent? Ajó et al. (2022) found that sample mass significantly impacted hormone concentrations and recommended that sample mass to solvent ratios should be kept consistent. If the sample measurements did vary, I suggest conducting an additional study to determine if sample mass had an impact on the hormone results.

Ajó, A.F., Hunt, K.E., Dillon, D., Uhart, M., Sironi, M., Rowntree, V. and Buck, C.L., 2022. Optimizing hormone extraction protocols for whale baleen: tackling questions of solvent: sample ratio and variation. General and Comparative Endocrinology, 315, p.113828.

We conducted a followup study (see above) and trust that our limited revisions to the text adequately address the reviewer's concern.

Lines 525-528: Were the samples run in duplicate? If not, how was the precision of the measurements determined?

Lines 525-528 do not address measurement precision, so the focus of the reviewer's comment is unclear to us. We did not run samples in duplicate because the entire powder sample was used up in each analysis. However, in lines 542-543, we report the measurement precision for testosterone and specify that it is "based on quality control solutions run with each batch." To address the reviewer's comment, we have modified the wording of this sentence. Pooled serum measurements are included in the data spreadsheet. Revised text reads:

New lines 569-570: "Measurement precision for testosterone was ± 3 % (standard error) based on 23 "Pooled Serum" quality control samples that were processed with batches of dentin samples."

EXTENDED DATA

Line 734: (Figure 6) How were the authors able to link the growth layers in each slab? Did the authors have possession of the entire tusk or just the slabs?

See response to Line 258 comment above. As detailed in the main text **(New lines 214-216)**, *this specimen consists of the 10 discs only. The remainder of the tusk was retained (and we suspect, eventually sold) by the diamond company that had collected it. However, we do have measurements of the lengths of each of the 10 segments of the tusk (cut in the warehouse of the diamond company) before our slabs were cut from each one. Extended Data Figure 6 shows the alignment of records in the 10 discs. Feature correlations were accomplished through qualitative comparison of features visible in thin sections of adjacent slabs. In response to the comment for line 258, we added text to the figure caption to clarify how corresponding features in different discs were identified.*

SUPPLEMENTARY DISCUSSION

Lines 842-843: Please include reference.

New line 879: *Here we added a citation for the already referenced article by Hall-Martin (2010), Role of musth in the reproductive strategy of the African elephant.*

Lines 878-883: The pattern observed in testosterone data could also be due to differences in sample mass, if the mass of the samples or the sample to solvent ratio were inconsistent.

This has been addressed in detail above. We did a series of analyses to make sure that our analyses do not have a mass effect that could have impacted our results. We also made a few moderate revisions to address this concern.

Lines 891-893: The authors are only able to conclude that testosterone measured in dentin is reflective of the animal's hormone concentrations, since the results followed a pattern expected if an animal was experiencing an annual musth. The other three hormones have not been validated for their biological relevance.

Because the manuscript focuses on changes in testosterone, we understand the reviewer's comment. However, it was not only testosterone that reflected musth episodes, but androstenedione as well (with less dramatic changes). In fact, our observation that testosterone and its precursor androstenedione both show a pattern while progesterone and cortisol do not (this holds true for both the modern and fossil male tusks) is further support for these patterns being 'real.' However, to respond to the reviewer's recommendation, we have modified the text to clarify by replacing "steroid hormones detected in each of the three…" with "testosterone and androstenedione, which generally tracks testosterone in each of the three…"

The new sentence reads:

New lines 927-929: “Based on these considerations, we conclude that testosterone and androstenedione, which generally tracks testosterone in each of the three specimens, provide interpretable reflections of the animal’s hormone concentrations during its life.”

Other changes:

New lines 27-28: *changed* “circulating concentrations at a precise time” *to* “immediate circulating concentrations” *for readability. Changed* “whereas hair permits an assessment of steroids” *to* “whereas keratinous tissues permit assessments of steroids” *to make sense of the added citation for baleen here (as recommended by Reviewer #2).*

New lines 33-34: *Changed* “high-resolution” *to* “fine-scale” *and further modified the sentence, removing* “to measure testosterone, progesterone, androstenedione, and cortisol” *to eliminate the implication that we targeted those four steroids (in fact, these are just the four that were present of the 20+ that we looked for).* The new sentence reads:

“Here we use liquid chromatography tandem mass spectrometry (LC-MS/MS) paired with fine-scale serial sampling to measure steroid hormones in modern and fossil tusk dentin.”

New line 45: *removed unnecessary word: “various”*

New line 46: *For clarification, changed* “Based on the sample mass required and analytical precision” *to* “Based on the small sample mass required for analytical precision”

New line 52: *To correct unclear wording, swapped the order of “close relationship to Asian elephants” and “abundant well-preserved remains”. Sentence now reads:*

“Many aspects of woolly mammoth reproductive physiology and behaviour are unknown despite their abundant well-preserved remains and close relationship to Asian elephants (*Elephas maximus*).”

New lines 62-63: *For brevity, removed “that would have been” and “various.” New sentence reads:*
“On the other hand, measurements of steroid hormones could provide direct evidence of physiological changes linked to reproductive behaviours.”

New line 88: *added “also” for increased clarity*

New line 90: *To enhance clarity, changed “measurements” to “steroid concentrations”*

New line 93: *“5 to 100 mg” was changed to “1 to 100 mg” because although none of the tusk samples reported were less than 10 mg, we included 1 mg samples in various tests during method development.*

New line 105: *For brevity, removed "the testosterone precursor" before "androstenedione"*

New line 105: *Replaced "and" with "to". Text now reads:* "... the ratio of testosterone to androstenedione in serum…"

New line 119: *hyphen added in "life-history stages"*

New line 123: *to clarify the units, we have changed "are reported in… pg mg^{-1} for tusk samples" to "are reported in… pg per mg dentin for tusk samples". We leave "pg ml^{-1} for serum samples" because we consider that to be conventional.*

New lines 128-131: *Revised sentences to reflect that museum specimen 118186 actually only consists of one of the tusks and one of the feet. Also removed the euphemistic "collected" in order to be more straightforward with our treatment of the specimen's history and acquisition. The new text reads:*

"UMMP 118186 includes the right tusks and preserved soles of the a feet foot of from an adult male African elephant (*Loxodonta africana*) that was collected killed near Maun, Botswana, between late September and early October of 1963. The right tusk , which was sampled for this study, measures 202 cm along its outside curve, has a 143 mm diameter at its proximal end, and has a 60 cm deep pulp cavity."

New line 137: *to improve clarity, replaced "during the dry season (July-October in Botswana generally, …" with "during Botswana's dry season (July-October, …"*

New line 141: *removed redundant word, "sporadically"*

New line 142: *to improve specificity, replaced "Supplementary Information" with "Supplementary Notes"*

New line 145: *added citation to Smith and Fisher, 2011 (referenced elsewhere already) as the basis for linking the "deep pulp cavity" to being an animal in its "prime."*

New line 150-151: *For clarification, changed "could reflect a transition from" to "could reflect short-lived musth episodes during the transition from"*

New line 158: for more precise wording, replaced "established previously in that same year" with "obtained for earlier in that same year"

New line 162: *Figure 1 revised with y-axis label changed from "Testosterone (pg mg^{-1})" to "Dentin testosterone (pg mg^{-1})"*

New line 162: *Figure 1 secondary axis label changed from "* $^{18}O_{VSMOW}$ *" changed to "* $^{18}O_{Carb}$ *(‰ VSMOW)"*

New line 162: *Figure 1 revised to remove musth marker (Ⓜ) from minor testosterone peaks in 1957 and 1958 years.*

New line 166: *to improve specificity, replaced "Supplementary Information" with "Supplementary Notes"*

New line 173: *To improve readability, replaced "are consistent with ^{18}O data for seasonality" with "are consistent with ^{18}O evidence of seasonality"*

New lines 164-185: *corrected the figure panel lettering in the caption (changed from capital A-E to lowercase a-e)*

New lines 170-171: *added "from dentin carbonate" to clarify the source of the ^{18}O values.*

New line 171: *removed "at the end of the rainy season" and "at the end of the dry season" and moved this information to the caption for figure panel e (the precipitation plot). The new text for e reads:*
(**New lines 179-181:)** "e, Precipitation records for Maun, Botswana for the years 1962 and 1963 relate to the ^{18}O fluctuations (d) with oxygen being relatively enriched at the end of the dry season and depleted at the end of the rainy season."

New line 174-175: *To replace "Abbreviations:..."* **(removed from New line 185)**, *added "Musth episodes interpreted from surges in testosterone are marked with Ⓜ."*

New lines 182-185: *Revised wording for panel 'f' for improved clarity. New text mirrors added text for Fig. 2 caption and reads:*

"f, LC-MS/MS plots of integration peaks for testosterone (solid line) and the deuterated testosterone internal standard that was added to each sample (dashed line) for four sequential samples that show the increase in the final musth episode (See Extended Data Figures 1, 2)."

New line 188: *To improve clarity, replaced "specimen" with "tusk"*

New line 190: *To improve readability, replaced "The specimen" with "It"*

New line 223: *To improve tone, changed "Cortisol again shows up" to "Cortisol was measurable"*

New line 225: *For brevity, "one of the samples" was changed to "one sample"*

New line 246: *Figure 2 revised with y-axis labels changed from "Testosterone (pg mg^{-1})" to "Dentin testosterone (pg mg^{-1})"*

New line 258: *changed "in" to "on"*

New line 260: *To more clearly identify the symbol used to mark musth episodes, we replaced 'M' with Ⓜ. To improve readability, we replaced "testosterone increases" with "surges in testosterone"*

New line 261: *Corrected text that read "... included here as horizontal dashed lines in each plot". Now reads "... included here as a horizontal dashed line in each plot."*

New line 270: *Added missing comma after (~6.5 cm)*

New line 272: *To correct grammatical error, changed "consistent with it being" to "consistent with its being"*

New line 272: *Replaced "Meanwhile" with the more appropriate word in this instance, "Thus". It is because we know that this is not a young individual that we interpret the small diameter as evidence of it being female.*

New line 287: *Figure 3 revised with y-axis label changed from "Testosterone (pg mg^{-1})" to "Dentin testosterone (pg mg^{-1})"*

New line 300-301: *For clarity and to make consistent with other figure captions, added:*
"Testosterone measurements are reported in pg per mg of dentin."

New line 317: *Because we discuss the possibility of ion suppression as a source for lower values in the ancient mammoth dentin, we added* "or other technical factors associated with preservation" *after* "... might be a result of degradation"

New line 472-473: *Added clarification that procedures are described elsewhere citing already referenced article (Fisher et al., 2014):*
"Following previously described procedures, (Fisher et al., 2014) we planned sampling to produce serial data from tusk growth records."

New line 473: *Added missed comma after "For each sample slab"*

New line 502-503: *replaced "measured" with "observed in thin sections" because no measurements from thin sections are reported in this manuscript. Thin sections were in fact measured, but that data is not relevant here. The only contribution from thin sections was correlation of the 10 discs from UMMP 118711, and that was done qualitatively using observations of features in thin sections.*

New line 513: *reference to Extended Data Fig. 7 (showing the sample planning/milling process) was moved to the end of the previous sentence. This sentence documenting sample masses has been revised and sample masses are not reflected in Extended Data Fig. 7.*

New line 528: *added hyphen to make "5-mg" a compound modifier of "aliquots" and added "dentin" to modify "powder" here*

New line 531: *Added "Isotope" before "analyses" for clarity.*

New lines 531-533: *For increased clarity, replaced "aliquots" with "portions". New sentence reads:*

"Isotope analyses were performed in the University of Michigan Stable Isotope lab (UMSIL) where 1-mg portions of each pre-treated powder sample were loaded into a Finnigan MAT Kiel IV preparation device…"

New lines 537-579: *Various wording changes to Steroid analyses section for increased clarity.*

New text for the first few sentences reads:
New lines 538-542: "To extract steroid analytes from powder samples, we pipetted 1.5 ml 100 % HPLC-grade methanol and 100 µl internal standard made from deuterated forms of each analyte in 40 % methanol into the 2-ml centrifuge tubes containing powder samples. We then added ~250 mg Lysing Matrix D beads (MP Biomedicals) to each tube to prevent clumping during agitation and rotation."

Final sentence in second paragraph now reads:
New lines 567-568: "Steroid concentrations for tusk dentin samples were calculated by adjusting measured concentrations according to sample mass to produce results expressed in pg mg^{-1} dentin powder."

Third paragraph contains various minor changes and now reads:
New lines 569-579: "Measurement precision for testosterone was ± 3 % (standard error) based on 23 "Pooled Serum" quality control samples that were processed with batches of dentin samples. Although numerous tests showed that steroid extraction from tusk powders was negligible after the first two hours (Extended Data Fig. 8), exhaustive steroid extraction was not necessary for these experiments and was not conclusively established. To accommodate the uncertain and likely partial preservation of steroids in the fossil materials, these experiments were designed only to detect temporal changes in steroids preserved within each tusk record. Normalising testosterone records relative to series minimum (Extended Data Fig. 9, Supplementary Discussion) and relative to androstenedione measurements (Extended Data Fig. 10, Supplementary Discussion) may produce more comparable datasets from specimens with differential preservation, but for this study, the results are clear without normalisation."

New line 588: *To be more precise, we changed "lab managers" to "instrument operators"*

New line 827: *corrected verb number "were" to "was"*

New line 834: *replaced "extraction" with "recovery"*

New line 836: *replaced "appear to be" with "remained"*

New line 837: *replaced "might" with "could"*

New lines 859-875 (Data tables): *Various minor formatting issues resolved. This will ultimately be submitted as a single .xlsx file with multiple sheets.*

New line 923: *replaced "in" with "for" ("contributes to lower average values for the woolly mammoth tusk…"*

New lines 929-930: *Added sentence to explicitly state the conclusion that androstenedione is biologically relevant in these analyses. New text reads:*
"Adrostenedione generally tracks testosterone with increased concentrations during musth, and we conclude that it also provides a biologically meaningful endocrine record."

New lines 931-983: *Section titled "Dentin steroids" has undergone major revision to address some of the concerns of the two reviewers and to better document the significance of each steroid detected in this study. The revised section is split into 5 paragraphs. In the first 4, we document what we see in dentin for each of the steroids (in order, testosterone, androstenedione, progesterone, cortisol) and discuss its biological relevance. The final paragraph addresses the inconsistent measurements for cortisol and progesterone (contrasting with the interpretable data for testosterone and androstenedione) and offers a way of understanding what might be happening that leads to "underrepresentation" for cortisol, seemingly accurate (and in any case useful) representations for testosterone and androstenedione, and then "overrepresentation" for progesterone. The new text is copied in full here:*

"**Dentin steroids**

Testosterone was quantitable in nearly all samples and preserved records of clear transitions to and from peak concentrations during musth. Also, except for when samples were apparently affected by heat (Extended Data Fig. 9), testosterone results were replicated consistently. Of the four steroids we detected in tusk dentin samples, testosterone produced the most robust, reliable data. The fact that the pattern in the male mammoth record is consistent with the one recorded in the modern elephant tusk implies that the changes in testosterone composition were not produced from groundwater contamination or other diagenetic effects.

Androstenedione was consistent, repeatable, and tracked loosely with testosterone. Considering all tusk samples, measurements of androstenedione were moderately correlated with testosterone (the Pearson correlation coefficient for paired results was 0.51 with $p < .001$, $n = 240$) and were generally higher during musth episodes. The mean of androstenedione values during musth episodes (including three sequential samples that form the tip of each testosterone peak) was about 120 % of the mean value for the between-musth intervals (including three sequential samples that form the trough in testosterone between each pair of musth peaks). However, androstenedione in most samples was lower than testosterone and the musth-related fluctuations in the androstenedione record were far less pronounced and clear

than those recorded in testosterone. Small amounts of androstenedione come from the testes, but most of the testicular androstenedione is converted to testosterone and therefore does not contribute to systemic levels in the blood. Androstenedione is also produced by the adrenal, and similar values in the male and female records suggest that the androstenedione in tusk dentin is primarily of adrenal origin. A recent study shows that cortisol increases during musth in Asian elephants, (Glaeser et al., 2022) and adrenal-derived androstenedione would be expected to rise in parallel. Thus, we cannot distinguish whether the small increases in androstenedione in the tusks during musth are of testicular, adrenal, or combined origin.

Progesterone had the highest mean values in dentin analyses. Although it is possible that it would be useful for investigating female reproductive patterns (progesterone approximately doubles during early pregnancy in elephant serum), progesterone was highly variable in our analyses. A possible explanation for the consistent abundance of progesterone is that progesterone is the most hydrophobic and last to elute in our method, which will be discussed below.

Cortisol recovery was highly inconsistent. Replicated results from duplicates show that high concentrations are not due to post-sampling contamination. Secondary inclusion in the dentin from contaminants on the cut surface of the tusk remains possible, even if unlikely due to careful handling and cleaning procedures employed during sampling. Perhaps more likely explanations for generally low cortisol responses are degradation due to oxidation (Davio et al., 2020) or bacterial degradation (Ridlon et al., 2013) over time. We have found that cortisol converts to 11-hydroxyandrostenedione during dehydration of extracted serum samples, especially when exposed to heat. However, these explanations fail to clearly explain the sporadic, seemingly random high values.

The source of this variability in progesterone and cortisol is unclear, but might be related to the chemical properties of dentin and those of each steroid. Cortisol is the most hydrophilic of the four, eluting near the beginning of the chromatographic sequence as methanol concentration is ramped up and thus could be the least likely to bind to dentin (or remain bound after initial incorporation). Progesterone, on the other hand, is the most hydrophobic of the four, eluting at the end of the chromatographic sequence and seems to show a particular affinity to binding to dentin (relative to the other steroids measured, it is much higher in dentin than in serum). One possible explanation for the highly variable progesterone concentrations is that it is a consequence of small changes in dentin structure and composition throughout each year that are reflected in differential binding of progesterone. On the other hand, testosterone and androstenedione are intermediate, eluting in the middle of the chromatographic sequence and appear to reflect systemic concentrations in blood. If this explanation is correct, it would imply that cortisol and progesterone in dentin are unreliable, and that testosterone and androstenedione are within the range of hydrophobicity required to provide biologically relevant records."

New line 1020: *replaced "which also shows up clearly" with "which is also consistently measured"*

New line 1055: *Added section title, "Steroid survey"*

New lines 1055-1064: *removed capitalization for all steroid names in list and replaced all "OH" with "hydroxy"*

New line 1058: *Removed the abbreviation "-AD" from the listing for "Androstenedione."*

New line 1059: *spelled out "Estriol" and removed abbreviation "E3"*

New line 1062: *inserted alpha character for 16-alpha-hydroxyprogesterone and inserted beta character for 11-beta-hydroxyprogesterone*

New line 1063: *spelled out "deoxycorticosterone" and removed abbreviation "DOC"*

New lines 1065-1075: *To address* Nature*'s requirement for a* ***Data availability statement****, we have added the section, "Data availability" to the Supplementary Notes. It contains a listing of DOIs for the CT scans of specimens that are posted on the UM library's site "Deep Blue Data"*

New line 1081: *References for Bronk Ramsey, 2009 and Reimer et al., 2020 converted to appropriate format and are now cited with numbers.*

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The authors are to be congratulated for the care and attention they have paid to revising their manuscript. I remain excited about the elegance of the work and look forward to seeing the paper published. I have annotated the manuscript with some suggestions for what are mostly editorial changes rather scientific changes. Two changes I really would like to see made are (i) reconsideration of the use of the term "aliquot" - 5 mg is not an aliquot whereas one fifth is - an aliquot refers to a specific fraction of a larger whole. (ii) the use of the word sample remains confusing throughout the manuscript and SI. Call them what they are, i.e. dentine powder, extracts, etc. As an analytical chemist of long experience this make papers so much easier to read. Sample is conveniently used in laboratory discussions but in writing the excessive use of the word sample becomes confusing. Sorry for being a pedant about these two points. I do not need to see the paper again.

Referee #2 (Remarks to the Author):

Comments for the author
I would like to congratulate the author for their improvements to the manuscript and for taking the time to address the points raised by the reviewers. I would also like to applaud the authors for taking the additional steps of conducting a controlled test to investigate sample mass effects. I have no further comments or concerns.

Author Rebuttals to First Revision:

Second reviews

Referee #1 (Remarks to the Author):

The authors are to be congratulated for the care and attention they have paid to revising their manuscript. I remain excited about the elegance of the work and look forward to seeing the paper published. I have annotated the manuscript with some suggestions for what are mostly editorial changes rather scientific changes. Two changes I really would like to see made are (i) reconsideration of the use of the term "aliquot" - 5 mg is not an aliquot whereas one fifth is - an aliquot refers to a specific fraction of a larger whole. (ii) the use of the word sample remains confusing throughout the manuscript and SI. Call them what they are, i.e. dentine powder, extracts, etc. As an analytical chemist of long experience this make papers so much easier to read. Sample is conveniently used in laboratory discussions but in writing the excessive use of the word sample becomes confusing. Sorry for being a pedant about these two points. I do not need to see the paper again.

Once again, we greatly appreciate the referee's additional feedback and careful scrutiny of our text. In our final submission, we have accepted nearly all of the suggestions. In particular, we now better understand the referee's request regarding the word "aliquot" and have made the requested change. We have also done our best to minimize use of "sample". Specific comments were copied from the manuscript pdf annotated by the referee and pasted below with our line-by-line responses.

Line 36: tooth-associated may be a better terminolgy and as bound suggests chemical bonding when this may not exist - it maybe purely an entrappment.

We agree with the referee's assessment and accepted the suggestion to replace "bound" *with* "associated" *here and below*.

Line 45: tissues maybe a better word

We have accepted the referee's suggestion and replaced "materials" *with* "tissues".

Lines 48-49: could add veterinary

We have accepted the referee's suggestion and added "veterinary" *to the list.*

Line 55: change to palaeobologists

We thank the referee for catching this error and have corrected the spelling.

Lines 262-266: no explanation of what the light and dark greens correspond to.

In this document, where word count is not such an issue, we can describe our plotting convention in more detail. Our plots of integration peaks for testosterone (solid line) and its internal standard dashed line) are distinguished only by the character of their bounding curves. Both peaks are highlighted by the same translucent green, avoiding the appearance of a different hue in areas of overlap, had we used a two-colour convention. We have added the following text to the legends for Figures 1 and 2:

"Integration peaks are shaded with translucent green; overlapping portions appear darker."

Line 327: I'm not sure there is any evidence for the steroids being bound - associated yes but not chemically bound - just entrapped.

We accepted the referee's suggestion and changed "bound" *to* "associated".

Line 508: no need for the word samples - Dentine powder is enough.

We accepted the referee's suggestion and changed "were" *to* "was".

Line 513: Dentine powder

We accepted the referee's suggestion here, but with the spelling "dentin" that we consistently use throughout.

Line 517: dentine powder

We accepted the referee's suggestion here, but with the spelling "dentin" that we consistently use throughout.

Line 523: ?

Line 524: ? call them what they are - every time I see the word sample written I have ask myself what is this? so call them what they are!

To address the referee suggestions for Lines 523-524 we dropped the word "samples" *in the first instance, modified the verb tense as subsequently required, and modified the end of the sentence to eliminate the use of* "samples" *there as well. In addition, we replaced the word* "selected" *with the more appropriate word* "obtained."

The new text reads: "Elephant serum was obtained from the Toledo Zoo & Aquarium repository of surpluses from previous veterinary analyses."

Line 528: I was taught, and the dictionary definition is, that "an aliquot is a proportion of a larger whole, especially when a sample is taken for chemical analysis" e.g. a one fifth aliquot. Please change this.

Replaced "aliquots" *here and in line 529 with* "portions". *Thanks for returning to this. We understand the referee's point better now than after the first review cycle.*

Line 538: dentine powders - no need for the word samples

We accepted the referee's suggestion here, but with the spelling "dentin" that we consistently use throughout.

Line 549: So what is this? It's a different sample to the powder so the use of the same word is very confusing - call it what it is - an extract?

To address the referee's concern here, we have replaced "Samples…" *with* "Reconstituted extracts…"

Line 554: Now we're talking about extracts?

Here, we removed the word "samples" *and restructured the sentence to be in active voice.*

The new text reads: "In each batch, we included multiple controls (processed methanol, processed water, an internal pooled serum, and a quality control mixed from stock steroids) and at least four calibration standards with concentrations spanning the range detected in tusks."

Line 567: superfluois word

We accepted the referee's recommendation here and removed "samples."

Line 568: dentine powder

We accepted the referee's recommendation here (using alternate spelling, "dentin"*) AND removed the word* "powder" *from the end of the sentence, which now reads:* "... results expressed in pg mg^{-1} dentin."

Line 570: confusing use of samples?

To improve clarity, we replaced "dentin samples" *with* "dentin powders" *and removed the first instance of* "samples."

The new text reads: "Measurement precision for testosterone was ± 3 % (standard error) based on results from 23 "pooled serum" quality controls that were processed with batches of dentin powders."

Line 968: sera

To accommodate the referee's request to reduce the use of "samples," *we accepted the recommendation here. We replaced* "serum samples" *with* "sera".

Lines 1012-1013: use different words than samples - call them what they are. Sample block = tusk block?

For the two uses here, we accepted "tusk block" *as a more precise alternative to* "sample block."

Line 1035: tusk powder?

Here, we think the word "sample" *can be eliminated from* "sample mass", so we have removed it. *We also revised the earlier use of samples (to which this mass refers) to clarify that we are referring to baleen and faeces from previous publications.*

The new text reads: "Nevertheless, in published results, analyses that used more than 10 mg of faeces or baleen produced consistent results independent of mass, and all of our dentin powders were above 10 mg (all but two were greater than 20 mg)."

Line 1036: ?

Line 1038: ?

To improve the clarity in Lines 1036 and 1038, we revised the sentence to remove ambiguous uses of "sample." *We also added* "bulk sample" *in reference to the vial of dentin powder that was collected specifically to test for replication of results with different masses, removed the misleading word,*

"routinely", *removed the superfluous* "pg ml-1"*, and reordered portions of the sentence for readability.*

The new text reads: "In our tests using a series of weighed portions of a homogenised bulk sample of dentin powder, masses as low as 1 mg produced clean signals of testosterone, androstenedione, and progesterone (cortisol was inconsistent), but the results from portions of less than 5 mg were mostly below concentrations of the lowest calibration standard and therefore could not be confidently quantitated."

Line 1040: ?

Replaced "Some of the smallest samples" *with* "The lowest masses"

Line 1043: ?

Replaced "sample masses" *with* "dentin powder masses"

Line 1044: ?

Replaced "of results in samples more massive than" *with* "in results for masses greater than"

Line 1046: ?

Replaced "sample masses" *with* "dentin powder masses"

Line 1047: ?

Removed "sample"

New text reads: "Because masses tended to be…"

Line 1048 (x2): ?

To be consistent with previous references, we replaced "sample slab" *with* "tusk block".

We also replaced the second instance of "sample masses" *with* "powder masses."

Referee #2 (Remarks to the Author):

Comments for the author

I would like to congratulate the author for their improvements to the manuscript and for taking the time to address the points raised by the reviewers. I would also like to applaud the authors for taking the additional steps of conducting a controlled test to investigate sample mass effects. I have no further comments or concerns.

We are gratetul for this referee's positive response.