

Semicircular canals shed light on bottleneck events in the evolution of the Neanderthal clade.

Corresponding Author: Dr Alessandro Urciuoli

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The manuscript „Semicircular canals shed light on bottleneck events in the evolution of the Neanderthal clade” is an interesting contribution to discussing the possible evolutionary mechanisms shaping the morphological variation among Middle and Late Pleistocene humans. However, before considering it for publication, there are some conceptual and methodological issues that could be improved and/or expanded.

1- INTRODUCTION: The authors mention that they will “test for the presence of genetic bottleneck events”, however, I do not think this is the precise word that they mean to use. Why not morphological bottlenecks? I understand that the authors want to reconstruct population history through the study of the bony labyrinth, and that this structure showed to be a good proxy for that, but from the moment in which you study morphological variation, you won't be able to predict genetic variation, simply because they are affected by different factors. I understand that you are aiming to study random processes or genetic drift by using the bony labyrinth, so I suggest choosing carefully your words through the manuscript.

2- INTRODUCTION: On the same line, the author says: “Thus, changes in the underlying genotype should generally be reflected in the phenotype, and greater phenotypic variation should be present before a drastic reduction of the genetic variation, such as a bottleneck. Thus, periods of drastic decrease in phenotypic variability would represent strong evidence for the occurrence of genetic bottlenecks”. Not necessarily all genetic changes have a correlate on the phenotype and the other way around. I understand the methodological approach of using the bony labyrinth to study random factors in evolution, but the authors should be careful with the way they are presenting this. Please revise the sentences in page 5, lines 103-107.

3- INTRODUCTION: The use of the word “variability” is also not correct, and I think the authors here mean “variation” instead. Variability means the tendency of a system to generate differences, while variation is the set of observed differences. See Hallgrímsson & Hall (2005) for clarification: Hallgrímsson, B. and Hall, B.K., 2005. Variation and variability: central concepts in biology. In Variation (pp. 1-7). Academic Press.

4- INTRODUCTION: Similarly, the use of the concept “population dynamics” is very vague. It usually implies the study of changes in space and time. So, for example, when the authors say “In this scenario, population dynamics would be the main driver of the Neanderthal clade evolution...” What do they mean exactly? Which is the evolutionary mechanism/factor they are referring to? Please check the use of this concept through the text and replace it for more specific concepts.

5- INTRODUCTION: It is unclear if the bony labyrinth of the Aroeira 3 specimen is plesiomorphic or not. The authors first say that is more primitive than the Sima de los Huesos humans, but also they said it resembles modern humans and early Homo. Could you please clarify these ideas, please?

6- INTRODUCTION/RESULTS: The paper would be much improved if the authors analyze and compare the variation in the cochlea vs the one described in the semicircular canals, or better justify why they are not studying the cochlea or the whole bony labyrinth to address the phylogenetic changes they are interested in. For instance, some works in carnivorans have shown that the cochlea is phylogenetically more informative, while the semicircular canals are influenced by ecological variables (see Schwab et al., 2019: Schwab, J.A., Kriwet, J., Weber, G.W. and Pfaff, C., 2019. Carnivoran hunting style and phylogeny reflected in bony labyrinth morphometry. Scientific Reports, 9(1), p.70.)

7- RESULTS: When the authors say that RMH show a tighter clustering than the whole Neanderthal clade, the reference to the Figure/Table that indicated that is missing.

8- RESULTS: It seems that the Krapina group presents the highest magnitude of variation of the whole sample, while the recent modern humans are the less diverse. Which could be the morphological/evolutionary reasons of that?

9- RESULTS: How do the authors explain the high variation of the Homo sapiens sample across PC3/PC4?

10- RESULTS: There are a lot of differences between PC1/PC2 and PC3/PC4, I would suggest the authors to add into the

Supplementary Material other combination of PCs such as PC1/PC3, PC2/PC3 for getting a better sense of the affinities presented by the different groups.

11- RESULTS: In addition to the variance and disparity indexes, the authors should provide the Euclidean distances in a separate table.

12- DISCUSSION: The authors should also consider the thicker canals of Neanderthals could be an evolutionary trade-off.

13- DISCUSSION: In page 12, lines 282-284, this sentence needs clarification/expansion. First, it is a strong claim saying that the samples SH is a single Middle Pleistocene population. Maybe they could say it was studied here as if it represents that. Also, they say that "phenotypic variability was partly recovered..." The use of the word variability is wrong here, and what does it mean by "recovered"?

14- DISCUSSION: The authors conclude that the results obtained provide strong support for Scenario 3, but how the variation in Homo sapiens could be explained within this scenario? Also, I don't think this is a straightforward result when looking at the PCS, which are the main results provided in the article. Maybe the Euclidean distances are showing that, but it is unclear if differences are significant.

15- DISCUSSION: In lines 335-339, the authors say that "the presence of distinctive and primitive features in early members of the Neanderthal lineage suggests mosaic evolution". Does this mean that Neanderthals also have derive features? Are those derived features shared with Homo sapiens or are they distinctive?

16- DISCUSSION: How do the authors interpret the lack of changes in terms of centroid size? This is an unexpected result, considering that even among human bony labyrinth there are significant differences in size.

17- MATERIALS AND METHODS: The authors say that the Homo sapiens sample is from Europe, but their associated chronology is unclear. For moments it says "recent" and in other passages "fossils". I request that the authors add that information on geographic location and chronology to Supplementary Table 1, but also, they should mention it on the text. In Supplementary Table 1, they should add a column specifying the population or site for each sample. Also, I recommend them to use the full names instead of abbreviations for the groupings in that table.

18- MATERIAL AND METHODS: Are the samples mentioned as "Universidad de Burgos" published? Please add a reference or more information.

19- MATERIAL AND METHODS: It would be relevant to add information on the sex of the samples too, which is not mentioned at all on the text. Some papers have found sexual dimorphism to affect the patterns of morphological variation, so how did the authors deal with this?

20- MATERIALS AND METHODS: Group dispersion, variance, and disparity. It is unclear which is the statistical significance of the results provided in Table 1. Also, these tests should be better explained and justified within the text.

21- FIGURE 1: The recent modern humans are not displayed; I would suggest adding them.

22- FIGURE 4: What does frequency mean in this context? This should be explained

Reviewer #2

(Remarks to the Author)

This study examines variation in the semicircular canals of the bony labyrinth among European hominins from the Middle Pleistocene to the present day. The authors use these morphological data as a proxy to infer patterns of genetic variation through time, which is a reasonable extrapolation based on published research, and which is important given the limited genetic data available for this time period as the authors explain. The authors use advanced and appropriate methods of statistical shape analysis to compare the overall shapes (and variability thereof) of these semicircular canals, and find that changes in shape variation corroborate other evidence for a population bottleneck in Neandertals between the Middle and Late Pleistocene. This is an important contribution to our understanding of human and Neandertal evolution.

The paper overall is very well written. The introduction clearly lays out the problem to be addressed, and provides a succinct summary of background information with a good balance between detail and brevity. The intro highlights that even though we have good genetic evidence for the Neandertals generally, there's not much earlier than 100,000 years ago, hence the need for a morphological proxy. I think the authors also make a good case for the somewhat limited Middle Pleistocene sampling (Sima, Krapina), making for a simple but elegant test of the bottleneck hypothesis. I love the figures: they are clean, relatively simple, and highly readable. The 3D PCA plot is pretty great, too. I also greatly appreciate the detailed description of the steps taken to collect and process the semicircular canals, which will ensure repeatability and allow others to add new data in the future to complement the data/analyses presented here.

I have one fairly big comment or concern, but hopefully the authors can address it without much difficulty or hassle. The main result of the study is that the decrease in variability between Krapina and later Neandertals supports the scenario of a population bottleneck in Neandertal evolution. However, looking at the Supplementary Table 1, most of the Krapina CT scans have the lowest resolution of nearly the entire sample in the analysis. Given the small absolute size of the bony labyrinth, the effects of scan resolution could be quite meaningful. For instance, Gunz (2012, J. Anat. 220 pp529–543) showed subtle but nontrivial differences between reconstructed 3D surfaces derived from micro vs. medical CT resolutions (Figures 5, 6, and 10 therein). Specifically, he compared labyrinth reconstructions from an original 0.09 mm³ voxel size to a down-sampled size of 0.30 mm³, which I think is slightly narrower than the full range of resolutions used in the present study.

Could this difference in scan resolution affect the results? Within the Krapina sample, Kr38.20 appears to have the highest resolution and falls at the positive PC1-2 shape extremes of Krapina in Fig. 3a, and is one of the only 2 Krapina specimens within the range of PC2 scores for the rest of the study sample. Similarly, Sima Cr6 has a much lower resolution than the rest of the Sima individuals, and it also plots at the extreme of the distribution (in fact the highest PC1 score in the analysis).

These may be coincidences but they do raise the question of whether/how scan resolution impacts the results, especially since the DMorph method compares whole surfaces.

So, the authors should explain why the unique aspects of Krapina labyrinths (MS lines 266–268 and Fig. 3a), the larger average centroid size within Krapina, and especially the greater variability within Krapina, are all “real” phenomena and not artefacts of their lower scan resolution. I don’t know if a small analysis (similar to Gunz 2012) or something as simple as examining the correlation between scan resolution and PC scores is necessary. I’m not as familiar with the DMorph method and related literature so perhaps there is already a paper about the issue of scan resolution that you could cite here.

Minor comments/concerns

I think the authors should add a little more information or explanation of how variability is measured, as this is mentioned only briefly in the main text and methods section. Specifically, what is ‘disparity,’ how does it relate to the other metrics (group dispersion and morphological variance), and why is it necessary to show us all three metrics in Table 1 and Figure 4 (i.e., do they tell us different things)? I’m not sure what to make of the result that dispersion is not significantly different between groups (MS lines 225-226) but variance and disparity do differ between groups (all pairwise group comparisons?) based on the bootstrap analysis. This could all be quickly and easily clarified within the main text, or possibly just in the “Disparity analysis” section (starting at MS line 422)

I could also use a little more information about how sample size might affect measures of variability. In the Methods it is stated that variability/disparity is bootstrapped to account for sample differences, but some more details here would be helpful, especially since Figure 4a-b (which is lovely) shows great overlap in bootstrap distributions between groups, at least to my eyes. I found the manual for the dispRity package which explains rarefaction is possible in the ‘dispRity’ function — was rarefaction performed (cf. Andrea Cardini’s research using landmark-based morphometrics)? My sense just eyeballing the data is that maybe this is not a problem, but empirically demonstrating this would be another important contribution that I don’t think would take much work (e.g. a rarefaction analysis of the larger modern human sample showing the smallest sample sizes that reliably return the same measures of variability as the full sample, cf. Andrea Cardini’s publications).

Minutiae

In the Supplementary Table 1, Tabun C1 and Ehringsdorf H3 are listed among the Middle Pleistocene sample, but it is not clear if these latter two were actually included as they do not seem to appear anywhere else in the paper or figures.

In Fig. 3, a point is labeled HA 4253 which I believe is Cro Magnon but is not specified as such within the documents that I can see. This could be quickly and easily noted in the Figure 3 caption (its plot symbol also appears different), and/or Supplementary Table 1.

How does variability of the labyrinth compare with variability of other systems? E.g. Krapina has the highest measurement of within-sample variability for the labyrinth compared to the other groups—have other studies of teeth, jaws, postcrania, etc. found a similar result?

I don’t love the abbreviations for fossil groups within the text, i.e., LN for late Neandertals and Kr for Krapina. Krapina is already short, so it probably doesn’t need to be abbreviated. The group “LN” in the paper appears to be abbreviated as simply “e” in Table 2 and as “cNE” in the Supplementary Table 1. So at the very least the name for this group should be consistent across the manuscript and other files. Later neandertals could be “LP” for Late Pleistocene, which I think is already an established abbreviation whereas LN is a common shorthand for ‘natural logarithm.’

Very minor

Line 55 replace “development” with “evolution” or something similar since these are related but distinct phenomena (many researchers and public discourse use the terms interchangeably, but this distinction is the small hill I will die on)

Line 80 replace the comma “,” at the very beginning of the line with a period “.”

The last sentence in the paragraph introducing the labyrinth (lines 111–114, “Housed within...”) is informative but seems out of place — I think it works well between the two sentences in line 108 (i.e., right before “Not only are petrosals...”, this way it anticipates “petrosals” in that sentence).

Line 117 remove “by” and the comma after “sagittal plane”

Line 292 replace “disposal” with “rejection”

Table 1: Why does Krapina’s dispersion have an ‘equals’ sign before the number?

I do like this study and manuscript a lot, though hopefully the scanning resolution question is not too consequential. Great work!

-Zach Cofran

Reviewer #3

(Remarks to the Author)

This paper analyses the surface morphology of the semicircular canals of the inner ear in Middle-Late Pleistocene hominins from western Eurasia to test competing demographic scenarios of the Neanderthal lineage. It does so by applying digital imaging and 3D geometric morphometrics techniques, as well as multivariate statistics (PCA) and analyses of morphological variance and disparity. The results are interpreted to support a model in which the Neanderthal lineage experiences a Late Pleistocene bottleneck, although further considerations are necessary to render this inference robust. While this study is of interest to the paleoanthropology community, several revisions must be made to make it accessible to the broader readership of this journal and, more importantly, some important analyses of measurement error must be conducted to ensure that the results are well founded. In addition, the theoretical background for assessing phylogeny, morphological disparity, and inferring a bottleneck event requires further elaboration at the outset (in the Introduction), with critical evaluation thereafter (in the Discussion), particularly with respect to evolutionary migration (i.e. gene flow/admixture) as a confounding factor. The points below compile suggestions for revisions to the manuscript.

Major points

1. Hominin phylogenetics: (a) Discussion of the phylogenetic relationship of Pleistocene hominins should be clarified. In particular, briefly specify in the second paragraph of the Introduction how the phylogenetic scenarios relate to diverse possible taxa (e.g. not only *Homo heidelbergensis* but also *H. rhodesiensis/bodoensis*, *erectus*, etc.; see e.g. Harvati & Reyes-Centeno 2022 and references therein) and/or paleo-demes (sensu Howell 1996, 1999). Clarifying the competing phylogenetic scenarios will then make it easier to refer to the bony labyrinth configuration across time. For example, the 5th paragraph of the Introduction (lines 115-137) should make explicit how the ancestral configuration is determined. (b) Clarify also the use of "lineage," "clade," and "paleodeme" throughout the manuscript: are they interchangeable? Importantly, (c) be explicit about how the phylogenetic scenarios justify the study design, particularly for the analysis of disparity (see also point 2, below). Finally, (d) consideration of admixture between lineages is necessary and the confounding effects that it has on the study design, particularly with respect to the inferences on bottleneck events. In particular, the evidence for gene flow from the sapiens lineage to Neanderthals is now robust (e.g. Kuhlwilm et al. 2016, cited in the study; see also Posth et al. 2017, this journal). How does migration affect morphology and, ultimately, the inferences made from the study's analyses? Is it possible to separate out the effects of admixture and bottleneck events?
2. Evolutionary Theory, Variation and Disparity: Associated to clarifying the phylogenetic context of the study, a brief introduction to the concept of disparity is necessary at the outset in order to justify such analysis. Aside from the fact that the concept might be unfamiliar to a general readership, it is often confounded with the concept of variation and diversity. There is a rich theoretical history to explain patterns of disparity, primarily along ecological and developmental hypotheses and it is unclear how this study is approaching it (i.e. how it is justifying the analysis). After providing the concept's background in the Introduction, the Discussion should then refer back to that underlying theory in the context of the results. Note that for some of the readership, an analysis of morphological disparity would typically entail quantification of distance between species. In this case, the calculation is between hypothetical paleo-demes (and a recent modern human group with one fossil), so the study design needs to be properly justified in this context (see also point 3, below, on sampling strategy).
3. Sampling: while access to data might preclude the inclusion of other fossils, some discussion is necessary on how/whether doing so could change results. What other fossils could be included? What are the considerations of the geographical and temporal sampling (i.e. what spatiotemporal "gaps" need to be filled?)? For the analysis on disparity and the inferences drawn therein, what are the implications of not sampling a hypothetical ancestor of the Neanderthal and sapiens lineage (i.e. *heidelbergensis/rhodesiensis/bodoensis* and/or *ergaster/erectus*)? Similarly, how might the inclusion of early Middle-Late Pleistocene anatomically modern humans and later Late Pleistocene modern humans inform analyses on disparity and variance (i.e. arguably two separate demes)? What is the effect of only including one fossil in the recent modern human group? Is this justified because the entire Cro-Magnon fossil assemblage is assumed to be recent (see Fu et al. 2013 for radiocarbon date, <https://doi.org/10.1016/j.cub.2013.02.044>), because the date range between specimens is relatively small compared to the possible range in the other groups, or for another reason?
4. Methodology: because the diffeomorphic morphometry (DMorph) method relies on the comparison of surfaces, a calculation of measurement error must be presented for (a) the segmentation procedure used to develop the surfaces and (b) for differences between micro-CT and medical CT data. The former (a) is straightforward and would consist of segmenting a bony labyrinth multiple (3+) times to assess any surface variation within the same individual, noting also variation in thresholding parameters or other procedures (e.g. manual delineation of structures). The latter (b) would consist of segmenting the bony labyrinth for an individual where both micro-CT and medical CT data is available and assessing surface variation between the two. While both error tests are important, the latter is of particular concern given that medical CT data comprises voxel sizes that are non-isometric. For relatively large anatomical elements (e.g. a cranium, where a relatively high number of voxels are represented), measurement error will likely be negligible; however, for relatively small anatomical elements (e.g. the bony labyrinth, where a relatively low number of voxels are represented), measurement error might be significant. Non-isometric medical CT scans will inevitably create a bias along the Cartesian direction comprising the larger voxel size, with variation added if the specimens are scanned in different directions. Therefore, it is important to show whether the combination of micro-CT and medical CT data affects how the surface is quantified when applying DMorph. This should be tested in one individual at minimum but ideally in multiple individuals in order to show that variation associated with voxel type is minimal and relatively less than the variation otherwise observed per hominin group sampled. In other words, the variation associated with data type (micro-CT vs medical CT) should be less than the variation from any of the hominin groups. Note also that the degree of error can have an impact on the visualization of deformations along PCs. For example, surfaces based on non-isometric voxels might appear "...more voluminous and inflated..." (line 178) than surfaces based on isometric voxels.

Minor points:

5. Methods: clarify the mesh [pre]alignment process (lines 384-390). While unclear, I understood this to mean that all sampled surfaces are superimposed manually rather than in a standardized fashion that relies on, for example, homologous landmarks. It is not encouraging to read that the meshes were "...roughly prealigned..." so an error test would also be in

order here if the pre-alignment process is not standardized.

6. Methods: justify the PC selection process. Was choosing PCs explaining ~90% of variance an arbitrary cut-off, those explaining >1% variance, or something else? See, e.g. Reyes-Centeno, Harvati, and Jäger 2016 (and references therein) for example of systematic method for PC selection (<https://doi.org/10.1038/srep36645>).

7. Results, PCA: when discussing the results, (a) be specific about how the overlap of groups is determined (i.e. based on the convex hull) and avoid using “discriminate” since this is not a linear discriminate or canonical variate analysis here. Put another way, what would the results look like if you visualize 95% confidence intervals instead of convex hulls? It is fine to visualize with only convex hulls, but then the description and interpretation of results should be specific to these analyses. In addition, (b) quantify how clusters can be compared. For example, I don’t see the “tight(er) cluster[ing]” (lines 214, 221) that is interpreted for the RMH and LN groups. I might be convinced if the area (in the 2D case of 2 PCs) or volume (in the 3D case of 3 PCs) of the convex hulls is quantified and compared. Otherwise, it would suffice to refer to the group dispersion results (but see below).

8. Results, group dispersion: specify how significance in group dispersion values between two groups was determined, i.e. what statistical test was used to quantify significance.

9. Line 111: rather than highlighting the bony labyrinth as a “...marker of population dispersal in modern humans,” it would be better in this context to justify its use as “...population history of modern humans” to highlight its use in intra-specific variation.

10. Lines 118-119: amend to “...sagittal labyrinthine index, which quantifies how the lateral anal plane intersects the posterior canal superiorly (Fig. 3d-f)” (referring to the Figure will be useful for non-specialists, see also comment 30, below, on figure labels).

11. Line 121: change “...is not present in...” to “...is not present to the same degree...” (or similar amendment)

12. Lines 123-124: change “...Aroeira 3...SH humans and is also from the Iberian Peninsula,” to “...Aroeira 3 from Portugal...SH humans from Spain are geographically proximate within the Iberian Peninsula,” (or similar amendment) to provide reference to non-specialists

13. Line 125: remove “much” or otherwise quantify descriptor

14. Line 126: change “displaced” to “placed” (the former might connote a pathology or functional trait that reduces fitness)

15. Line 126: change “instead” to “thus” to provide continuity with previous sentence. Note also that some readers might otherwise mischaracterize the use of “primitive” (see also point 1, above, on phylogenetics)

16. Line 367: change “...430 ka” to “...~430 ka”

17. Line 133: change “...this study...” to “...that study...” (to clarify the preceding reference rather than the present original study)

18. Lines 144-147: amend incomplete sentence (e.g. “Two geochronologically constrained populations...were sampled to maximize...”)

19. Lines 154, 155, 166: amend change equal symbol to approximately equal symbol (e.g. line 154: “SH = Kr = LN” to “SH ≈ Kr ≈ LN”)

20. Line 208: consider changing “...hiatus between RMH and the Neanderthal clade...” to “...space between the convex hulls formed by RMHs and Neanderthals...” Note that the term “hiatus” is awkward in this context and might be interpreted to imply an associated hiatus in the archaeological record.

21. Lines 218-220: disambiguate this sentence by removing the parenthetical information and including it so that the comparisons are properly understood.

22. Lines 380-381 and Table S1: Specify in Table S1 whether the right side of the bony labyrinth was used and make reference to it in Lines 380-381. In Lines 380-381, quantify the percentage of cases where mirroring was applied.

23. Line 414: clarify what the “DeterministicAtlas option” refers to: is this the sample-average template that enables the calculation of momenta deformation vectors?

24. Line 416: clarify what the raw shape data is (vector lengths?)

25. Line 425: briefly explain how disparity is calculated (i.e. how PC scores are treated).

26. Line 678: delete “Resume of the...” (or change to “Résumé of the...”/“Summary of the...”)

27. Line 685: delete “Three-dimensional...”

28. Lines 690-691: change “fossil modern human” to “recent modern human” (see also point 3, above, on sampling strategy)

29. Table 2: add label descriptor for Cro-Magnon 1

30. Figure 3: add labels to one of the canal figures (e.g. Fig. 3c) which conveys the names of the canals. This will aid a non-specialist reader in understanding the morphology.

31. Figure 3 Legend, Lines 685-691: clarify “HA” label (Cro-Magnon)

--

Hugo Reyes-Centeno
Assistant Professor, University of Kentucky

Department of Anthropology
211 Lafferty Hall, Lexington, KY 40506 USA
William S. Webb Museum of Anthropology
1020 Export St, Lexington, KY 40504 USA
heva.as.uky.edu | educelab.engr.uky.edu

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The new version of the manuscript, "Semicircular canals shed light on bottleneck events in the evolution of the Neanderthal clade" addresses most of my and the other reviewer's comments appropriately. Moreover, it stands out as a much-improved version of the manuscript. I would like to thank the authors for their great work. I am convinced that this is an excellent contribution to Nature Communications, and I would like to congratulate the authors for their great job. However, I still have some minor concerns, which I believe won't be difficult for the authors to reconsider and resolve at this stage. I hope these points help strengthen some aspects of the manuscript that still need attention. All my comments refer to the PDF version with tracked changes.

1. I agree with the authors that the morphological variation of the bony labyrinth is an excellent proxy for discussing random factors in evolution and, as such, random genetic variation for reconstructing population or evolutionary history. I appreciate that the authors have improved many parts of the text following my previous comments on this. However, the way this has been phrased in some passages of the manuscript is still ambiguous. In general terms, whenever they mention "genetic bottleneck," I would be more cautious and replace it with "bottleneck" (please check throughout the manuscript). This is mainly because not all genetic variation is random. This is perfectly expressed, for example, in lines 223-225 of the manuscript. Therefore, I suggest changing the wording, particularly in these cases: 1) In the abstract, line 35, instead of "genetic variation," it should say something like random or neutral factors; 2) In the abstract, line 38, instead of "genetic variation," it should say something like evolutionary history/population history; 3) In the Introduction, line 188, instead of "genetic variation," change to random or neutral variation.
2. In relation to chronology, I think it would be clearer for the readers, especially considering the wide audience of Nature Communications, to include, in brackets, the equivalent chronology to the MIS stage whenever possible as they are mentioned throughout the text, particularly in the Introduction and Discussion. I know that this information is available in Figure 1, but personally, I think it will make the narrative clearer.
3. Introduction, lines 130-131: The authors argue that "periods of drastic decrease in disparity might represent strong evidence for the occurrence of genetic bottlenecks." The first question that comes to my mind is: A decrease in disparity means an increase in affinities among the groups analyzed, so why couldn't this be the result of other processes, for example, gene flow or shared/parallel adaptations?
4. Discussion, lines 401-403: The authors argue that "the increased phenotypic diversity of Krapina...might be the result of neutral phenotypic evolution..." It is not clear to me why the authors suggest neutral evolution in this context, which is usually the result of drift or mutations. Usually, in small and isolated populations, drift is associated with a decrease in genetic variation, while mutations happen at such a slow rate that they would not be sufficient to explain such changes in diversity. I would instead suggest other factors to explain this, like the ones suggested in the following sentences (lines 404-416: natural selection, gene flow, phenotypic plasticity). Unless I misunderstood and when they refer to neutral evolution meaning drift, they imply the fixation of unique traits within a population, increasing the diversity between the compared groups. If this is the case, this clarification should be added to the manuscript.
5. Discussion, line 482: The authors mention "discarded variation," which I think sounds a bit odd. I would replace it with retained/relict/remnant variation, if I understand correctly what they are trying to convey here.

Minor comments:

- Line 61: change to "sometimes regarded as a separate species..", changing the position of the "a"
Line 495: change to "the evidence provided by semicircular canal shape variation", deleting the "and"
Line 499: I think the authors mean "more rapidly/fast" instead of "more frequently"

Lumila Paula Menéndez, University of Vienna

Reviewer #2

(Remarks to the Author)

The authors have gone to great lengths to address my and the other reviewers' comments and concerns on the first draft of the manuscript (NCOMMS-24-28932). I agree with and am convinced by all the responses and decisions relating to the comments raised by myself and the other reviewers.

I greatly appreciate the in-depth analysis and responses to my original questions about scan resolution as well (and apologies that I had not already seen your previous work that addressed this issue).

I have no major concerns about this new draft, and only have a few minor comments or suggestions which are detailed below.

Altogether, I thought the original manuscript was quite strong and the revised manuscript is even stronger, so I will be happy to see this paper in print soon. I have learned a lot about the vestibular system (as well as later Pleistocene human evolution and analytical methods) from this paper, thank you!

Minor comments

Abstract, line 35 of revised text: change "drop" to "reduction"

Line 81 of revised text: change "leading" to "led"

Line 137: add "s" to other "Among others..."

Line 137: “clockwise” should be replaced or specified as it will have a different meaning for the right vs. left sides. E.g. Spoor et al. (2003, p. 158) specified “When viewing the lateral aspect of the left temporal bone and labyrinth this morphology can be seen as a joint clockwise rotation....” Along these lines, perhaps Figure 1 should specify that images depict the left vestibular system.

Line 245: the word “modern” is missing the ‘n’

Line 299: there is a large space between “customarily ... considered”

Line 307: The sentence “Our results further remark on the post-Sima....” I think it would be clearer if it was stated or reviewed exactly what the results tell. I.e., “Our results show that a late Neandertal-like morphology evolved...” highlighting how this anatomy appeared/changed between Sima, Krapina, and the final late Neandertal sample. (this explanation does appear later in lines 395–397)

Line 316: Change “conception of” to “argument for.” Note also that “drastic” appears in two sentences in lines 316 and 317, so stylistically one of these instances could be changed to similar word like “extreme” or “severe.”

Line 340: Change “Tecent” to “Recent”

Line 364: remove the comma (,) between “variation also eliminates”

Reviewer #3

(Remarks to the Author)

The revisions have greatly improved this manuscript and properly addressed comments and suggestions from the earlier draft. Some minor points of revision are enumerated below.

1. Methods, sexing: Briefly explain the biological sexing of the recent human samples (e.g. karyotyping, assigned-at-birth, self-reported, etc.) and cite the source for the fossil samples (e.g. with superscript references in in Table S1), noting limitations. If any of the fossils were assigned a sex for this study, explain the criteria used and its limitations. Consider adopting the term “sexual differences” rather than “sexual dimorphism” for this general science journal (see, e.g., review by Khramtsova et al. 2019. doi:10.1038/s41576-018-0083-1).

2. Within the text, clarify the PC selection process. This was explained in the rebuttal, but not explained in the main text.

3. Abstract, lines 35-38: this sentence has a tone of speculation, so consider rephrasing this sentence for clarity and making an explicit hypothesis instead, e.g. “The goal of this study was to quantify the degree of phenotypic disparity in the Neanderthal clade. We hypothesized that a reduction in bony labyrinth disparity is indicative of the underlying genetic variation resulting from bottleneck events.”

4. Abstract, line 42: change “...providing strong evidence for...” to “...supporting our hypothesis...”

5. Citation edits: superscript citations should not be accompanied by additional text. If additional text is necessary, then include it parenthetically as a normal script and the citation number preceded by “Ref.” For example, line 66 would be: “... (see Ref. 24 and references therein for a review). In most/all instances, the accompanying text can be excluded and the citations can simply be in superscript.

6. Introduction, line 80: change to “...late Neanderthal morphology...” or “...late Neanderthals’ morphology...”

7. Introduction, line 82: remove reference to H. heidelbergensis since previous paragraph has established the possibility of H. antecessor as ancestral

8. Introduction, line 146: change “other” to “others”

9. Introduction, lines 164-167: amend for clarity, e.g. “...bony labyrinths highlighted...variation in the latter group.”

10. Throughout the manuscript, use site names of Sima de los Huesos and Krapina as an adjective for the sample/group/population rather than a noun, e.g.:

- a. Introduction, lines 191-194, 198-202, 203-207: amend for clarity, e.g. “...after the Sima de los Huesos population...”
- b. Results, lines 238-240: amend for clarity, e.g. “...in the Sima de los Huesos, Krapina, and late Neanderthal samples.”
- c. Results, line 256: amend for clarity, e.g. “...in the Krapina sample relative to the late Neanderthals sample.”
- d. Results, line 302: amend for clarity, e.g. “The Krapina sample is more variable...”
- e. etc.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The manuscript „Semicircular canals shed light on bottleneck events in the evolution of the Neanderthal clade” is an interesting contribution to discussing the possible evolutionary mechanisms shaping the morphological variation among Middle and Late Pleistocene humans. However, before considering it for publication, there are some conceptual and methodological issues that could be improved and/or expanded.

We thank the Reviewer #1 for their positive comments about the manuscript. All the points they raised were carefully considered and addressed as specified in the following point-by-point response.

1- INTRODUCTION: The authors mention that they will “test for the presence of genetic bottleneck events”, however, I do not think this is the precise word that they mean to use. Why not morphological bottlenecks? I understand that the authors want to reconstruct population history through the study of the bony labyrinth, and that this structure showed to be a good proxy for that, but from the moment in which you study morphological variation, you won't be able to predict genetic variation, simply because they are affected by different factors. I understand that you are aiming to study random processes or genetic drift by using the bony labyrinth, so I suggest choosing carefully your words through the manuscript.

We agree with the Reviewer that caution must be used when considering the relationship between phenotype and genotype. We thus have rephrased the sentence of the abstract to better convey the hypothetical nature of our statement: “The goal of this study was to test for the presence of drops in bony labyrinth disparity during the evolution of the Neanderthal clade, possibly hinting for bottleneck events in the underlying genetic variation”. We would like to point out that several analyses showed that there might be a close relationship between phenetic and genetic variation when the previous underwent to neutral evolutionary processes, as shown in several studies (e.g., Harvati and Weaver, 2006; Manica et al., 2007; Roseman and Weaver, 2007; Reyes-Centeno et al., 2014). This is precisely the case of bony labyrinth shape, which shows a phylogenetic signal that suggests that it evolved neutrally both in primates (Urciuoli et al., 2020) and other mammalian groups (Mennecart et al., 2023), and it has been already used to successfully track dispersals in modern humans (Ponce de León et al., 2018) and has highlighted the presence of variance patterns that match those found in the underlying genetic variation (Stoessel et al., 2024). Hence, although we have somewhat toned down our statements in a few instances throughout the manuscript, we are overall reassured that our approach lays on a solid and well-accepted research foundations.

2- INTRODUCTION: On the same line, the author says: “Thus, changes in the underlying genotype should generally be reflected in the phenotype, and greater phenotypic variation should be present before a drastic reduction of the genetic variation, such as a bottleneck. Thus, periods of drastic decrease in phenotypic variability would represent strong evidence for the occurrence of genetic bottlenecks”. Not necessarily all genetic changes have a correlate on the phenotype and the other way around. I understand the methodological approach of using the bony labyrinth to study random factors in evolution, but the authors should be careful with the way they are presenting this. Please revise the sentences in page 5, lines 103-107.

Following the same reasoning explained above, we have toned down our statement (now in lines 126-131 of the track-changes manuscript). We also included additional reasoning to justify why consider it is feasible to rely on structures that evolved neutrally to confidently infer the underlying genetic variation history.

3- INTRODUCTION: The use of the word “variability” is also not correct, and I think the authors here mean “variation” instead. Variability means the tendency of a system to generate differences, while variation is the set of observed differences. See Hallgrímsson & Hall (2005) for clarification: Hallgrímsson, B. and Hall, B.K., 2005. Variation and variability: central concepts in biology. In Variation (pp. 1-7). Academic Press.

We thank the Reviewer for the suggestion. The usage of the words “variability” and “variation” was revised throughout the text to better convey the message of our reasoning. More specifically, we favored “variation” or “diversity” over “variability” for most instances, besides when we were referring to the amount of variation available for the clade at a given time.

4- INTRODUCTION: Similarly, the use of the concept “population dynamics” is very vague. It usually implies the study of changes in space and time. So, for example, when the authors say “In this scenario, population dynamics would be the main driver of the Neanderthal clade evolution...” What do they mean exactly? Which is the evolutionary mechanism/factor they are referring to? Please check the use of this concept through the text and replace it for more specific concepts.

We thank the Reviewer for spotting this issue. Particularly, we identified three instances where additional explanations or rephrasing were needed, which were the following:

- Introduction: “In this scenario, population dynamics would be the main driver of the Neanderthal clade evolution, and it would be compatible with an accretion model where Neanderthal features become fixed in the population through time and became best expressed in the late Neanderthals” – We clarified that we referred “(i.e., drift and/or extinction)” of these populations as the drivers for Neanderthal clade evolution.
- Discussion: “our results further strengthen the conception of a strong role played by population dynamics in the evolution of Neanderthals” – was rephrased to: “our results further strengthen the conception of a strong role played by drastic fluctuations in population size in the evolution of Neanderthals”
- Discussion: “hints for a reconsidering of the scenario for the origin and evolution of the Neanderthal clade and remarks on the role played by population dynamics” we changed “dynamics” for “structure”.

5- INTRODUCTION: It is unclear if the bony labyrinth of the Aroeira 3 specimen is plesiomorphic or not. The authors first say that is more primitive than the Sima de los Huesos humans, but also they said it resembles modern humans and early *Homo*. Could you please clarify these ideas, please?

We thank the Reviewer for spotting this lack of clarity in explaining the evolutionary background of inner ear evolution. Aroeira 3 was described as plesiomorphic relative to Neanderthals by Conde-Valverde et al. (2018) based on the shape of the anterior canal (wide instead of narrow), the relative size of the lateral canal (small relative to the anterior and posterior ones, as in early *Homo* species) and the lack of a high sagittal labyrinthine index (i.e., the lateral canal roughly divides the posterior canal plane in two equally-sized portions). Modern humans are plesiomorphic in several characters relative to the derived Neanderthal condition (as posited by Quam et al., 2016), as they retain a morphology that is already found in *Homo erectus*/*Homo ergaster*. Regardless of the reasoning above, we have now included an additional table (Table 1) to report and clarify the polarity of for bony labyrinth character states as inferred at different evolutionary stages of hominin evolution by previous analyses.

6- INTRODUCTION/RESULTS: The paper would be much improved if the authors analyze and compare the variation in the cochlea vs the one described in the semicircular canals, or better justify why they are not studying the cochlea or the whole bony labyrinth to address the phylogenetic changes they are

interested in. For instance, some works in carnivorans have shown that the cochlea is phylogenetically more informative, while the semicircular canals are influenced by ecological variables (see Schwab et al., 2019: Schwab, J.A., Kriwet, J., Weber, G.W. and Pfaff, C., 2019. Carnivoran hunting style and phylogeny reflected in bony labyrinth morphometry. *Scientific Reports*, 9(1), p.70.)

We appreciate the Reviewer's suggestion, which is aimed at making our manuscript more impactful and concur with the Reviewer that it would be very interesting to compare the amount of variation in the cochlea and the semicircular canals. However, this goes beyond the scope of the present analysis for several reasons.

First, these structures absolute very different sensory functions (i.e., hearing and balance for the cochlea and vestibular system, respectively), thus is plausible to expect similarly different selective pressures acting on them and shaping their morphologies. We thus believe that the addition of cochlear shape would, unfortunately, confuse the patterns of shape variation and disparity rather than ameliorate the potential of the present paper.

Second, the presence of a variable number of turns in the cochlea that partly overlaps and merge at different angular depths poses challenges for analyzing cochlear shape with a diffeomorphic approach. While it might be possible to study shape variation in the cochlea by relying on standard landmark-based 3D geometric morphometrics, there are additional shortcomings. Particularly, previous analyses have highlighted that cochlear shape represents a challenge for landmarks and semilandmarks too. This is due to the difficulties to define homologous landmarks (similarly to what happens with our diffeomorphic approach) along the cochlear course, which ultimately results in a so-called "horse-shoe effect" when the shape data is analyzed with a PCA (e.g., Beaudet et al., 2019; Veléz et al., 2023; del Río et al., 2023).

Third, even though the Reviewer is right about the results of Schwab et al. (2019), there are a few points that shall be taken into account. Schwab et al. (2019) computed the phylogenetic signal in linear measurements (i.e., the height and width of the canals). For most of the variables the authors did not recover a significant phylogenetic signal, while others shown a significant signal that was < 1 , thus implying that function might play a greater role than phylogeny in canal size evolution. This is not surprising, as it is known that semicircular canal size correlates with the locomotor proficiency (the so-called "agility score" in Spoor et al., 2007) of a species, where taxa with proportionally larger canals would be capable of engaging in swifter and more agile locomotor behaviors than species with smaller canals. In the present analysis, however, we did not rely on metrical variables and analyzed the overall semicircular canal and vestibule shape thanks to our diffeomorphic approach. This is a crucial point as canal shape is very poorly linked to the function of the embedded semicircular ducts when the structure is size-standardized. This is because size plays a fundamental role in assessing the overall proficiency of the vestibular system, so that very similarly shaped semicircular canals possess completely different functional properties if they vary in size (see David et al., 2022). Furthermore, even though the authors claim that the angle between the lateral semicircular canal and the cochlea (angl LSC/Coc in the cited paper) possesses a strong phylogenetic signal, they did not provide the results for this variable in the associated Supplementary Data 1 (where all other phylogenetic signal results are given), possibly due to an inadvertent omission.

Lastly, the inclusion of the cochlea for the sample here analyzed would limit the number of available individuals even further. This is due to the poor state of conservation of the cochlea in some of the Sima de los Huesos humans (e.g., Cr.5 and Cr.6), as well as the voxel size of some of the Krapina individuals (e.g., Krapina 39.13, 39.18, 39.3), the latter allowing the proper segmentation of the semicircular canals but not a fine visualization of cochlear details.

For the reasons above, we refrained to analyze the morphological disparity and variation of the cochlea in the present study.

7- RESULTS: When the authors say that RMH show a tighter clustering than the whole Neanderthal clade, the reference to the Figure/Table that indicated that is missing.

We thank the Reviewer for catching this. A reference to Figure 3 (showing the distribution along PC1-4) has been added.

8- RESULTS: It seems that the Krapina group presents the highest magnitude of variation of the whole sample, while the recent modern humans are the less diverse. Which could be the morphological/evolutionary reasons of that?

We agree with the Reviewer about the interpretation of our results, particularly regarding the high morphological diversity of Krapina. Although we cannot be entirely assertive about the reasons that led to the observed morphological disparity pattern, we now provide additional reasoning (current lines 398-413 of the track-changes version) in the Discussion section to discuss the possible causes of these different diversity levels.

9- RESULTS: How do the authors explain the high variation of the *Homo sapiens* sample across PC3/PC4?

The inclusion of an additional Upper Paleolithic individual in the analysis has caused a slight rotation of the morphospace. The consequence is that now modern humans appear very variable along PC4 only. This component appears to be most sensitive to changes occurring in lateral canal size. As noted by previous work, the genus *Homo* is generally characterized by the possession of larger anterior and posterior canals relative to the lateral one (e.g., Spoor et al., 1994), an adaptation that has been linked to the acquisition of bipedalism. However, this canal is one of the most variable portions of modern humans' vestibular system's bony encasing (and possibly even in anthropoids as a whole, see Urciuoli et al., 2020), even though the relative position of the lateral canal remains constant within the species. In our analysis, most individuals do show a small lateral canal relative to the vertical canals, yet just three individuals (namely, EMBR 121, EMBR 479, and F05) possess quite larger lateral canals and fall on marked negative scores for this reason. These are the main driver of the very broad range of modern humans for PC4, whereas most individuals. To clarify how to interpret the observed variation, we have now included a sentence about this point in the Results section.

10- RESULTS: There are lot of differences between PC1/PC2 and PC3/PC4, I would suggest the authors to add into the Supplementary Material other combination of PCs such as PC1/PC3, PC2/PC3 for getting a better sense of the affinities presented by the different groups.

We concur with the Reviewer regarding the presence of different patterns of shape variation along PC1-3, yet we respectfully disagree with them about the need to show a different combination of PCs in the bivariate scatterplots. While it would not be a problem for us to prepare additional figures/supplementary figures, these would be highly redundant with our 3D plot of the first three PCs (PC1-3), which is given exactly to address the problem specified by the Reviewer. Being an interactive HTML 3D plot, it allows the Reviewer and the readership to inspect the distribution of the group point clouds in total freedom and from any desired viewpoint.

11- RESULTS: In addition to the variance and disparity indexes, the authors should provide the Euclidean distances in a separate table.

The results to which Reviewer #1 is referring to no longer form part of the manuscript in its updated version. The group mean distances have been exchanged with pair-wise Euclidean distances computed for PC1-19, as explained more in detail in the response to one of the points raised by Reviewer #2. In this way, we can better characterize the distribution of the specimens within the occupied shape space, rather than just focusing on the distance between group averages. Note that the results for this additional metric are depicted in Figure 4, together with those of the other two metrics (individual-to-centroid distance and variance) we employed to measure morphological disparity. For this reason, we decided to maintain all the metrics in a single table (Table 2), as these should be analyzed together rather than one at a time.

12- DISCUSSION: The authors should also consider the thicker canals of Neanderthals could be an evolutionary trade-off.

Semicircular canal thickness is unfortunately a very poor predictor of the lumen size of the membranous ducts embedded in the bony canals (as demonstrated by Ramprasad et al., 1984). Even if we considered that enlarged canal lumen might reliably hint for a same enlargement of duct lumen, we are not entirely sure why this might be an evolutionary trade-off. Canal lumen size affects the sensitivity of the semicircular ducts, with larger ducts that are slower in detecting angular acceleration (e.g., Rabbitt et al., 2004). This is why we suggested a possible relationship between this newly reported feature and the different arrangement shown by Neanderthals in the head-neck compartment relative to modern humans. However, we are not aware if this was a drawback or an advantage for individuals belonging to the Neanderthal clade and thus it is unclear to us how it might represent an evolutionary trade-off. Furthermore, there is more than compelling evidence suggesting that Neanderthals were as proficient as we are in moving bipedally, which argues against a significant effect of canal lumen size on locomotion. We would also like to point out that modern humans share a similar canal thickness with early *Homo* species while these are thought to have been characterized by body proportions like those of Sima de los Huesos individuals (Arsuaga et al., 2015). Overall, we thus consider that canal lumen size is likely decoupled from body proportions and/or from adaptations related to locomotor behaviors and this is why we hinted that it might be related to non-adaptive processes such as genetic drift or pleiotropic effect (which have been already suggested as possible causes of evolutionary processes in the Neanderthal clade). We would like to finally remark that ongoing research by colleagues at the Natural History Museum of London is tackling the relationship between canal and duct lumen size, but that this is a very seminal stage. We thus hope that, together with the perfectioning of the available biomechanical models describing the functioning of the vestibular apparatus, it will be possible to address the Reviewer's comment in a more direct and thorough manner.

13- DISCUSSION: In page 12, lines 282-284, this sentence needs clarification/expansion. First, it is a strong claim saying that the samples SH is a single Middle Pleistocene population. Maybe they could say it was studied here as if it represents that. Also, they say that "phenotypic variability was partly recovered..." The use of the word variability is wrong here, and what does it mean by "recovered"?

Following the Reviewer's suggestion, we modified this period to ensure the clarity of our flow of thoughts. Particularly, we were less assertive regarding the relationship between the individuals of the Sima de los Huesos assemblage and thus modified the sentence as "an assemblage which likely represents a single Middle Pleistocene paleodeme", also providing a reference in support of our statement. Furthermore, we expanded our reasoning regarding the similarities between SH and modern human variability so that these "...suggest that the loss of variation caused by this bottleneck event was partly recovered as early as 430 ka⁷⁵, at least at the phenotype level". We used the verb "recover" in one of its main meanings that is "If something recovers from a period of weakness or difficulty, it improves or gets stronger again", as defined in the Collins dictionary. The dictionary also suggests that a synonym for "recover" is "bounce back", which is exactly the case of our sentence, where we imply that after an event of great variation loss

caused by a bottle neck, this (the variation) might have bounced back to (or close to) a pre-event diversity level.

14- DISCUSSION: The authors conclude that the results obtained provide strong support for Scenario 3, but how the variation in *Homo sapiens* could be explained within this scenario? Also, I don't think this is a straightforward result when looking at the PCS, which are the main results provided in the article. Maybe the Euclidean distances are showing that, but it is unclear if differences are significant.

We are grateful to the Reviewer for bringing our attention to a point where our reasoning was not clear enough. We would like to point out that the PCA axes shown in Figure 3 are only part of the results that we obtained and are discussing in the current contribution. Most importantly, with the PCA we wanted to understand the main patterns of shape variation occurring in the data, which allowed us to describe previously unidentified differences in the semicircular canal shape of members of the Neanderthal clade, and between them and modern humans. These results also show that modern humans are either similarly or less variable than the other groups in the first three PCs (SH is the most variable for PC1 and PC3, whereas modern humans are the most variable group for PC2 but closely followed by late Neanderthals), while they display a great variation along PC4, thus implying a greater range of morphological variation in lateral canal size than previously highlighted.

However, the four PCs here inspected account for roughly half of the total variance (~55%) and thus we decided to rely on morphological disparity and variance analyses (and now also a pairwise distance analysis) in the sample, whose results are shown in Figure 3 and consider a greater amount (~87%) of the total variance. These analyses demonstrate a significantly lower morphological disparity and variance (as well as pairwise distance) in late Neanderthals relative to all other considered groups, which demonstrate that they are the least variable group and is the main result that allows us to strongly support Scenario 3. Please note that the results of the dispersion analysis (i.e., the Euclidean distances) were removed from the current version of the manuscript following the suggestion of Reviewer #2.

In view of the obtained results, our interpretation of the condition found in the analyzed groups still holds and we thus suggest that Scenario #3 is the most likely, where the drop in morphological disparity and variance occurs after the age of Krapina. The moderately low variation of modern humans (lower than in Krapina and comparable or lower than in Sima de los Huesos, but always greater than in late Neanderthals; Figure 3) fits well in this scenario, as the sample is composed by European modern humans (both recent and fossils), known to have experienced a marked drop in genetic variation in recent times. Please consider that the role of modern humans in this analysis is that of representing the comparative sample, as their genetic history is well known and thus may represent a reference to which we are comparing the remaining group. However, as also remarked to Reviewer #3, they do not represent the main focus of our contribution and reconstructing their evolutionary history, in a similar way to what we do here with the Neanderthal clade, would require the inclusion of several additional specimens that are, unfortunately, not accessible to us.

15- DISCUSSION: In lines 335-339, the authors say that "the presence of distinctive and primitive features in early members of the Neanderthal lineage suggests mosaic evolution". Does this mean that Neanderthals also have derived features? Are those derived features shared with *Homo sapiens* or are they distinctive?

Neanderthals are autapomorphic for several features that accumulated throughout their evolutionary history in several anatomical regions, including the bony labyrinth of the inner ear. These features were highlighted in the Introduction and some of them we discussed in the Discussion. However, following the Reviewer's comments here and above, we now included a new table (Table 1) to be clearer about the

polarity of bony labyrinth morphology along hominin evolution. The table derives from previously published results by Conde-Valverde et al., 2018 and Quam et al., 2016.

16- DISCUSSION: How do the authors interpret the lack of changes in terms of centroid size? This is an unexpected result, considering that even among human bony labyrinth there are significant differences in size.

The Reviewer raises an interesting point that was only marginally considered in the previous version of the manuscript. This is because we were willing to exclude size that the observed variation of semicircular canal shape might have been a byproduct of size changes (that is, that it was mainly an effect of allometry). The Reviewer is however right regarding the overall small size differences in semicircular canal and vestibule size. Indeed, when differences in group means are tested by means of a one-way analysis of variance, we do not recover significant differences among them ($p > 0.05$ in all cases when Bonferroni post-hoc comparison is performed). This result is somewhat unexpected, as pointed out by the Reviewer, since previous analyses by Uhl et al. (2020) identified differences in modern human population in several bony labyrinth metric variables and ratios. We have added a reference in the results section to acknowledge this piece of information. Thanks!

17- MATERIALS AND METHODS: The authors say that the *Homo sapiens* sample is from Europe, but their associated chronology is unclear. For moments it says “recent” and in other passages “fossils”. I request that the authors add that information on geographic location and chronology to Supplementary Table 1, but also, they should mention it on the text. In Supplementary Table 1, they should add a column specifying the population or site for each sample. Also, I recommend them to use the full names instead of abbreviations for the groupings in that table.

We agree with the Reviewer that the wording that we used in the previous version of the manuscript might have been somewhat unclear and we thus modified it. In the first place, we have removed group abbreviations along the text, as well as in the tables, table headings, figures, and figure captions, as suggested by the Reviewers. We then removed any mention to “fossil modern humans” and substituted it for a clearer explanation stating that our modern human comparative sample is composed by extant modern humans (specifying the provenance of the three main group composing the extant sample) and by two Upper Paleolithic individuals. Following the Reviewer’s suggestion, we also included a new column (together with several other columns) in the Supplementary Table 1 to specify the geographic provenance of each of the specimens.

18- MATERIAL AND METHODS: Are the samples mentioned as “Universidad de Burgos” published? Please add a reference or more information.

Following the previous suggestion by the Reviewer, we added data regarding the provenance of the comparative human sample. However, we would like to respectfully point out that the reference for the previous publication of the data from the Universidad de Burgos was already included in the previous version of the manuscript, ref. 68. We thus did not implement further changes.

19- MATERIAL AND METHODS: It would be relevant to add information on the sex of the samples too, which is not mentioned at all on the text. Some papers have found sexual dimorphism to affect the patterns of morphological variation, so how did the authors deal with this?

We thank the Reviewer for pointing this out, as there is previous literature showing that some of the bony labyrinth variables might be affected by sexual dimorphism (e.g., Osipov et al., 2013; Uhl et al., 2020; Braga et al., 2019), although please notice that debate persists on this topic (see Ward et al. 2020). For what concerns the present analysis, sex is unavailable for several of the individuals included in the

comparative sample due to anonymization of the scanned individuals. Similarly, the sex of several fossil specimens belonging to the Krapina assemblage is unknown, as is that of a few late Neanderthals (e.g., La Quina H5, Amud 7, Engis 2, La Ferrassie 8), and of four juvenile individuals of the Sima de los Huesos assemblage (Cr16, AT1907, AT643, and AT421). With the data available to us regarding the sex of the individuals, it is not possible to identify any sex-related pattern in the shape data, as shown in the Supplementary Figure 2. Besides discussing these results in the main text in view of the possible sexual-dimorphism-related origin of Krapina's morphological variation, we included a new column in the Supplementary Table 1 to provide the sex for the individuals when known, as well as the aforementioned supplementary figure.

20- MATERIALS AND METHODS: Group dispersion, variance, and disparity. It is unclear which is the statistical significance of the results provided in Table 1. Also, these tests should be better explained and justified within the text.

Following the suggestion of Reviewer #2, we removed the dispersion analysis from the tests performed for the present study and changed it for a pair-wise Euclidean distances analysis, which allows a better understanding of how individuals are scattered within the multidimensional space and how closely individuals are packed.

21- FIGURE 1: The recent modern humans are not displayed; I would suggest adding them.

We added the modern humans (as well as the newly included Laugerie-Basse specimen) to Figure 1. The approximate geographical range of each one of the included collections is now shown by semi-transparent orange ellipses.

22- FIGURE 4: What does frequency mean in this context? This should be explained

Following the Reviewer's suggestion, we have included an additional explanation to the meaning of frequencies in interpreting the plots shown in the caption of Figure 4.

Reviewer #2 (Remarks to the Author):

This study examines variation in the semicircular canals of the bony labyrinth among European hominins from the Middle Pleistocene to the present day. The authors use these morphological data as a proxy to infer patterns of genetic variation through time, which is a reasonable extrapolation based on published research, and which is important given the limited genetic data available for this time period as the authors explain. The authors use advanced and appropriate methods of statistical shape analysis to compare the overall shapes (and variability thereof) of these semicircular canals, and find that changes in shape variation corroborate other evidence for a population bottleneck in Neandertals between the Middle and Late Pleistocene. This is an important contribution to our understanding of human and Neandertal evolution.

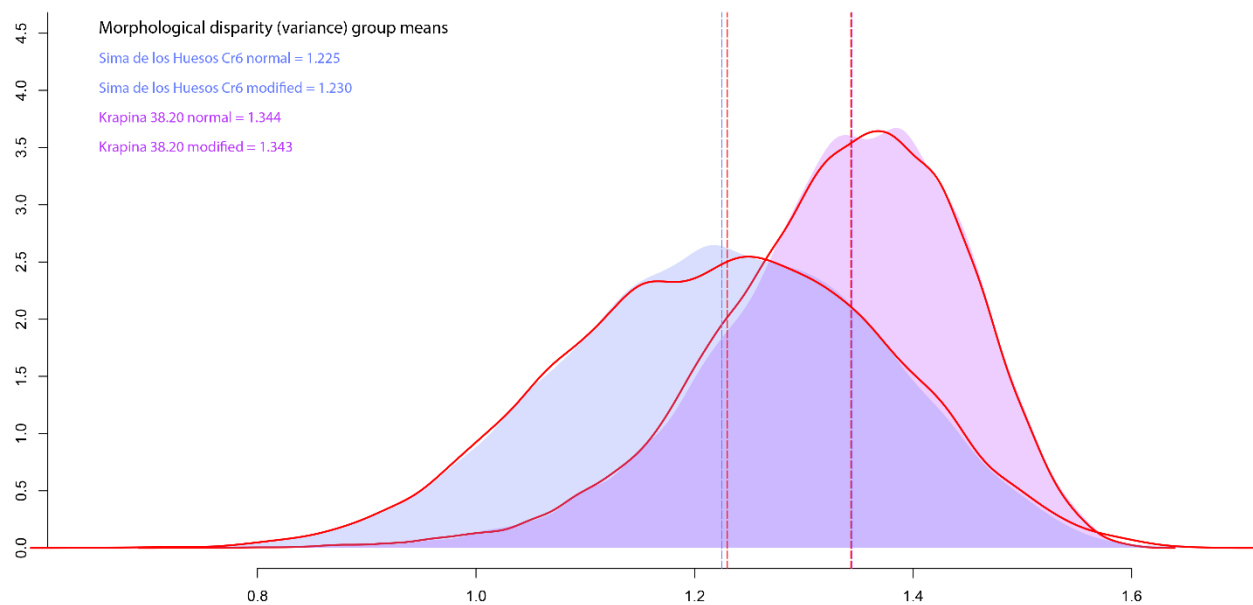
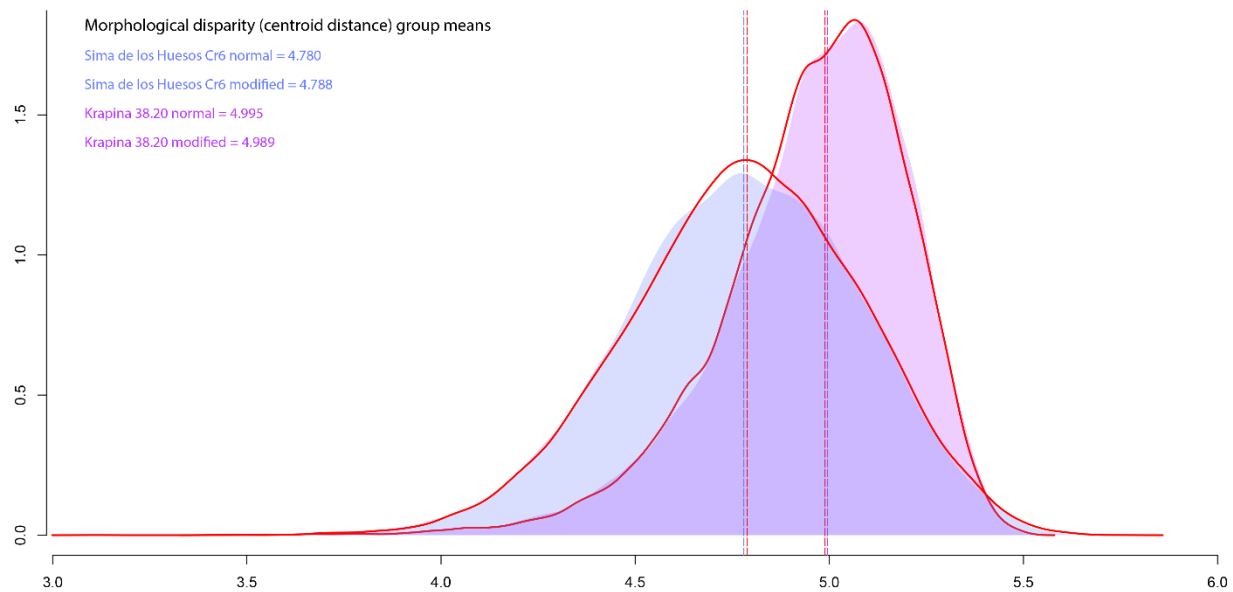
The paper overall is very well written. The introduction clearly lays out the problem to be addressed, and provides a succinct summary of background information with a good balance between detail and brevity. The intro highlights that even though we have good genetic evidence for the Neandertals generally, there's not much earlier than 100,000 years ago, hence the need for a morphological proxy. I think the authors also make a good case for the somewhat limited Middle Pleistocene sampling (Sima, Krapina), making for a simple but elegant test of the bottleneck hypothesis. I love the figures: they are clean, relatively simple, and highly readable. The 3D PCA plot is pretty great, too. I also greatly appreciate the detailed description of the steps taken to collect and process the semicircular canals, which will ensure repeatability and allow others to add new data in the future to complement the data/analyses presented here.

[We thank the Reviewer for the very positive comments about the manuscript!](#)

I have one fairly big comment or concern, but hopefully the authors can address it without much difficulty or hassle. The main result of the study is that the decrease in variability between Krapina and later Neandertals supports the scenario of a population bottleneck in Neandertal evolution. However, looking at the Supplementary Table 1, most of the Krapina CT scans have the lowest resolution of nearly the entire sample in the analysis. Given the small absolute size of the bony labyrinth, the effects of scan resolution could be quite meaningful. For instance, Gunz (2012, *J. Anat.* 220 pp529–543) showed subtle but nontrivial differences between reconstructed 3D surfaces derived from micro vs. medical CT resolutions (Figures 5, 6, and 10 therein). Specifically, he compared labyrinth reconstructions from an original 0.09 mm³ voxel size to a down-sampled size of 0.30 mm³, which I think is slightly narrower than the full range of resolutions used in the present study.

Could this difference in scan resolution affect the results? Within the Krapina sample, Kr38.20 appears to have the highest resolution and falls at the positive PC1-2 shape extremes of Krapina in Fig. 3a, and is one of the only 2 Krapina specimens within the range of PC2 scores for the rest of the study sample. Similarly, Sima Cr6 has a much lower resolution than the rest of the Sima individuals, and it also plots at the extreme of the distribution (in fact the highest PC1 score in the analysis). These may be coincidences but they do raise the question of whether/how scan resolution impacts the results, especially since the DMorph method compares whole surfaces.

[We thank the Reviewer for the constructive suggestion and very in-depth comments. Although we were already reassured by the results obtained by Urciuoli et al., 2021 \(ref. 59, supplementary online material \[SOM\]\) where the authors already demonstrated how DMorph is not very sensitive to changes in smoothing parameters and segmentation replicates, we added several tests to justify the inclusion of specimens with different voxel sizes, following the requests of both Reviewers #2 and #3. The results of these tests are given in Supplementary Figures 3-4 and show how scan parameters are not significantly affecting the results of our research. Regarding the two individuals identified by the Reviewer #2,](#)



So, the authors should explain why the unique aspects of Krapina labyrinths (MS lines 266–268 and Fig. 3a), the larger average centroid size within Krapina, and especially the greater variability within Krapina, are all “real” phenomena and not artefacts of their lower scan resolution. I don’t know if a small analysis (similar to Gunz 2012) or something as simple as examining the correlation between scan resolution and PC scores is necessary. I’m not as familiar with the DMorph method and related literature so perhaps there is already a paper about the issue of scan resolution that you could cite here.

In the first place, we corrected one value for the Krapina sample voxel size (namely, that of Krapina 39.13 from 0.6 to the correct value of 0.4 mm interslice), which was incorrectly widening the voxel size range for this group, we apologize for the typo! In addition, although the Reviewer is correct about the larger size of Krapina semicircular canals, we would like to point out that they show the narrowest range among the analyzed groups and that most individuals of Sima de los Huesos and late Neanderthals, as well as fewer

modern humans, fall within the range of Krapina, whereas a few even exceed it. We also would like to point out that, by addressing one of the points raised by Reviewer #1, we demonstrated that even though the average log-cSize is higher in Krapina, the means of the groups do not differ significantly. Finally, we used the segmentations performed for the points raised by the Reviewers #2 and #3 to compute the centroid size and included the result in Supplementary Figure 3. This allowed us to show how closely the CT- and μ CT-derived surfaces plot together and thus demonstrate that this possible source of bias is not artificially increasing the variation of the groups.

Minor comments/concerns

I think the authors should add a little more information or explanation of how variability is measured, as this is mentioned only briefly in the main text and methods section. Specifically, what is 'disparity,' how does it relate to the other metrics (group dispersion and morphological variance), and why is it necessary to show us all three metrics in Table 1 and Figure 4 (i.e., do they tell us different things)? I'm not sure what to make of the result that dispersion is not significantly different between groups (MS lines 225-226) but variance and disparity do differ between groups (all pairwise group comparisons?) based on the bootstrap analysis. This could all be quickly and easily clarified within the main text, or possibly just in the "Disparity analysis" section (starting at MS line 422)

Following the Reviewer's comment, we have included additional information and reasoning regarding the analysis of morphological diversity (i.e., disparity) in the literature and for the present paper. Particularly, we included a new paragraph at the end of the Introduction (as also suggested by Reviewer #3) to specify how disparity was intended for testing the hypotheses of the present study, as well as regarding the general concept of disparity and its normal use in the literature. We also included additional information in the Materials and Methods section about the metrics we measured for assessing the amount of disparity in the Neanderthal clade groups (as well as modern humans), specifying the different meanings of each one of them. Lastly, we removed the dispersion analysis and the related results (previously shown in Table 1) as it was computed on the raw shape data. Although not always the shape data shall be ordinated (i.e., there is not automatic benefit in performing a dimensional reduction prior to disparity analysis), shape data deriving from high-density geometric morphometric analysis (as is the case of our DMorph analysis with 1680 control points) may lead to artifacts in the obtained disparity. This is because very small fluctuations at the local level (which can be clearly regarded as noise) may artificially increase the amount of disparity. This is especially relevant for the groups composed entirely of fossils, where sediment filling may frequently cause small irregularities. These, as demonstrated in our segmentation replicate analysis, do not affect the general patterns of shape variation but may obfuscate disparity differences (Phillip Gunz, personal communication).

I could also use a little more information about how sample size might affect measures of variability. In the Methods it is stated that variability/disparity is bootstrapped to account for sample differences, but some more details here would be helpful, especially since Figure 4a-b (which is lovely) shows great overlap in bootstrap distributions between groups, at least to my eyes. I found the manual for the *disprity* package which explains rarefaction is possible in the '*disprity*' function — was rarefaction performed (cf. Andrea Cardini's research using landmark-based morphometrics)? My sense just eyeballing the data is that maybe this is not a problem, but empirically demonstrating this would be another important contribution that I don't think would take much work (e.g. a rarefaction analysis of the larger modern human sample showing the smallest sample sizes that reliably return the same measures of variability as the full sample, cf. Andrea Cardini's publications).

We agree with the Reviewer that sample size is a factor that must be considered when morphological disparity is inspected. To address this legitimate concern, we repeated the disparity analysis including the

rarefaction term in the `dispRity()` function, as suggested by the Reviewer. The rarefaction was set to “min”, which implies that the analysis would take the smallest group size for the bootstrapping process (in our case, 9 individuals). The rarefaction results (shown in Supplementary Table 2), show negligible differences from the original full sample for the morphological disparity, thus enabling us to rely on the latter for the purposes of the present paper. Slightly larger differences are recorded when the morphological variance is considered. These differences notwithstanding, Krapina individuals always show the greatest disparity, followed by Sima and recent modern humans, with late Neanderthals that appear the least variable group. Furthermore, the Wilcoxon test performed on the results with rarefaction still shows a significant difference among group morphological variance, overall confirming the soundness of our results.

Minutiae

In the Supplementary Table 1, Tabun C1 and Ehringsdorf H3 are listed among the Middle Pleistocene sample, but it is not clear if these latter two were actually included as they do not seem to appear anywhere else in the paper or figures.

We thank the Reviewer for catching this. The specimens were inspected in previous versions of the manuscript but not included in the analysis due to problems related with CT scan parameters (Ehringsdorf) and doubtful age of the deposits (Tabun C1). We thus removed them from Supplementary Table 1.

In Fig. 3, a point is labeled HA 4253 which I believe is Cro Magnon but is not specified as such within the documents that I can see. This could be quickly and easily noted in the Figure 3 caption (its plot symbol also appears different), and/or Supplementary Table 1.

We are grateful to the Reviewer for catching this. He is also right about the correspondence of MNHN HA 4253 to Cro-Magnon 1. We changed the figure caption accordingly adding specific reference to the specimen. Similarly, we added the specimen ID to the Supplementary Table 1 (as well as in Table 2) to avoid misconceptions.

How does variability of the labyrinth compare with variability of other systems? E.g. Krapina has the highest measurement of within-sample variability for the labyrinth compared to the other groups—have other studies of teeth, jaws, postcrania, etc. found a similar result?

This is a very interesting point, yet unfortunately there are no other studies that have performed disparity analyses in Krapina, at the best of our knowledge. Within hominins, Gunz et al., 2009 only approached the matter of morphological variation with a rigorous and statistically sound method. However, we did find out that Smith 1980 inspected the morphological variation of Krapina temporal bones. Although focusing mostly on sexual-dimorphism-related features, it is interesting that he recovered a somewhat higher variation than in modern European *Homo sapiens*. We have thus decided to consider this factor and included a few sentences to explain how sex-related differences do not seem to play a role in explaining the increased morphological disparity of Krapina (Supplementary Figure 2).

I don't love the abbreviations for fossil groups within the text, i.e., LN for late Neanderthals and Kr for Krapina. Krapina is already short, so it probably doesn't need to be abbreviated. The group “LN” in the paper appears to be abbreviated as simply “e” in Table 2 and as “cNE” in the Supplementary Table 1. So at the very least the name for this group should be consistent across the manuscript and other files. Later neanderthals could be “LP” for Late Pleistocene, which I think is already an established abbreviation whereas LN is a common shorthand for ‘natural logarithm.’

We thank the Reviewer for pinpointing this issue. We mistakenly left a previous version of the abbreviation in the main text and supplementary material. Following the Reviewer's suggestion, we now spelled out the full name of each group to ensure clarity and ease the reading of our manuscript.

Very minor

Line 55 replace “development” with “evolution” or something similar since these are related but distinct phenomena (many researchers and public discourse use the terms interchangeably, but this distinction is the small hill I will die on)

The text was changed according to the Reviewer's request.

Line 80 replace the comma “,” at the very beginning of the line with a period “.”

Thanks! Done.

The last sentence in the paragraph introducing the labyrinth (lines 111–114, “Housed within...”) is informative but seems out of place — I think it works well between the two sentences in line 108 (i.e., right before “Not only are petrosals...”, this way it anticipates “petrosals” in that sentence).

We concur with the Reviewer about the position of the sentence and moved it following his suggestion.

Line 117 remove “by” and the comma after “sagittal plane”

The editing was included as suggested by the Reviewer.

Line 292 replace “disposal” with “rejection”

We agreed with the suggestion.

Table 1: Why does Krapina's dispersion have an ‘equals’ sign before the number?

Thanks for drawing our attention to this. We removed the ‘equals’ sign as it was a typographical error.

I do like this study and manuscript a lot, though hopefully the scanning resolution question is not too consequential. Great work!

-Zach Cofran

We are very grateful to Dr. Cofran for his detailed comments and very positive consideration.

Reviewer #3 (Remarks to the Author):

This paper analyses the surface morphology of the semicircular canals of the inner ear in Middle-Late Pleistocene hominins from western Eurasia to test competing demographic scenarios of the Neanderthal lineage. It does so by applying digital imaging and 3D geometric morphometrics techniques, as well as multivariate statistics (PCA) and analyses of morphological variance and disparity. The results are interpreted to support a model in which the Neanderthal lineage experiences a Late Pleistocene bottleneck, although further considerations are necessary to render this inference robust. While this study is of interest to the paleoanthropology community, several revisions must be made to make it accessible to the broader readership of this journal and, more importantly, some important analyses of measurement error must be conducted to ensure that the results are well founded. In addition, the theoretical background for assessing phylogeny, morphological disparity, and inferring a bottleneck event requires further elaboration at the outset (in the Introduction), with critical evaluation thereafter (in the Discussion), particularly with respect to evolutionary migration (i.e. gene flow/admixture) as a confounding factor.

Several changes were implemented in the updated version of the manuscript following the Reviewer's suggestions (see below). Particularly, we are now providing an expanded human evolution background, which allow us to better pinpoint the questions we want to address, as well as define more precisely the phylogenetic basis upon which we designed the four scenarios at test. The role of admixture events was also directly tackled in the Introduction and then discussed in light on the obtained results. We also included further explanation regarding disparity analyses and the way their results should be interpreted by dedicating an entire section of the Material and Methods to this point. Lastly, we more direct about the outlook of the present work and identified which individuals might be sampled in the future to address the several remaining questions about the evolution of the Neanderthal clade. We thus believe that the revised draft our paper might be more easily accessed by a public with a broad interest in scientific literature, such is that of *Nature Communications*.

The points below compile suggestions for revisions to the manuscript.

Major points

1. Hominin phylogenetics: (a) Discussion of the phylogenetic relationship of Pleistocene hominins should be clarified. In particular, briefly specify in the second paragraph of the Introduction how the phylogenetic scenarios relate to diverse possible taxa (e.g. not only *Homo heidelbergensis* but also *H. rhodesiensis/bodoensis*, *erectus*, etc.; see e.g. Harvati & Reyes-Centeno 2022 and references therein) and/or paleo-demes (sensu Howell 1996, 1999).

Following the Reviewer's suggestions, we improved the paragraph in the introduction where we lay the phylogenetic background for the present work. Particularly, we included additional references and expanded our reasoning regarding the ongoing debate about how to interpret the Middle Pleistocene fossil record.

Clarifying the competing phylogenetic scenarios will then make it easier to refer to the bony labyrinth configuration across time. For example, the 5th paragraph of the Introduction (lines 115-137) should make explicit how the ancestral configuration is determined.

To address Reviewer's point, as well as the comments of Reviewer #1, we now added a table (Table 1) to provide a recap of the ancestral state condition for several characters described for the bony labyrinth. These polarities are extracted from previously published works (Quam et al., 2016 and Conde-Valverde et al., 2018, for the semicircular canals and cochlea, respectively) and are based on the morphology known for early hominin taxa (mainly, australopiths), early *Homo* (*Homo ergaster*), and chimpanzees.

Furthermore, we included explanations regarding the retention of primitive traits by *Homo sapiens* to better clarify the phylogenetic context regarding the evolution of bony labyrinth shape in hominins.

(b) Clarify also the use of “lineage,” “clade,” and “paleodeme” throughout the manuscript: are they interchangeable?

We thank the Reviewer for drawing our attention to this possible source of misunderstanding, particularly in view of the persisting debate about hominin phylogeny. In the manuscript, we used “lineage” and “clade” in a rather interchangeable manner. Indeed, following the definition of these words as provided by the Oxford dictionary, a lineage represents: “*the direct descent from an ancestor; a sequence of species each of which is considered to have evolved from its predecessor*”. Similarly, a clade constitutes: “*A term derived from the Greek *klados*, meaning a twig or branch. In cladistics, or phylogenetic systematics, it refers to a lineage branch that results from splitting in an earlier lineage. A split produces two distinct new taxa, each of which is represented as a clade or branch in a phylogenetic diagram*”. Hence, we interchanged them with the goal of achieving a pleasing writing style for our manuscript and thus avoid the repetition and abuse of the word “clade” (which is the wording we will prefer over lineage, if the Reviewer does not concur with our reasoning).

Conversely, with paleodeme we refer to a past population of a species. Indeed, in all circumstances it was used to refer to one or more populations constituting the total population (or metapopulation) of a species at a given time in the past. Just in one case, we used the wording “stem paleodeme”. Our idea was that of referring to a deme that might not have kept contributing to the gene pool of the total population. However, since the wording might have been misunderstandable, we removed the word “stem” and kept only “with some features that might have evolved locally in paleodemes”.

Importantly, (c) be explicit about how the phylogenetic scenarios justify the study design, particularly for the analysis of disparity (see also point 2, below).

Although we are aware that phylogeny of Middle Pleistocene hominins is still far from being resolved, we consider that this issue is not affecting the analysis as it is currently structured. This is because our study is limited to the specimens postdating the Neanderthal-Denisovan split. This is surely the case of the Late Pleistocene individuals included in the analysis and the late Middle Pleistocene remains recovered from Krapina. Similarly, the analysis of nuclear aDNA of the Sima de los Huesos humans (Meyer et al., 2016; etc.) undoubtedly places this Middle Pleistocene group already within the Neandertal clade (irrespective of them representing a separate species or being considered as part of *Homo neanderthalensis*, conceived as a chronospecies), although in a much basal position. For this same reason we did not include the Portuguese specimen Aroeira 3, as the affinities of this individual are still a matter of debate, and we would be unable to confidently assign it to an a priori defined group within the clade.

Finally, (d) consideration of admixture between lineages is necessary and the confounding effects that it has on the study design, particularly with respect to the inferences on bottleneck events. In particular, the evidence for gene flow from the sapiens lineage to Neanderthals is now robust (e.g. Kuhlwilm et al. 2016, cited in the study; see also Posth et al. 2017, this journal). How does migration affect morphology and, ultimately, the inferences made from the study’s analyses? Is it possible to separate out the effects of admixture and bottleneck events?

We thank the Reviewer for his comment, and we agree with him on the fact that gene flow and admixture likely played a role in shaping the evolution of modern humans and Neanderthals (and in more in general, hominins). We have thus added a new paragraph where we discuss the possibility that admixture and hybridization might represent possible sources of the increased variation shown by Krapina. However, given that these phenomena are currently considered as sources of morphological and genetic diversity,

we kept our reasoning regarding the interpretation of the four scenarios we presented in the introduction. Indeed, Li et al. (2024) and Posth et al. (2017) highlight that Neanderthals show the signature of gene flow from modern humans during the Middle Pleistocene and, possibly, by the boundary of the Middle and Late Pleistocene (Li et al. 2024). If we assume that admixture is one of the main causes of the observed morphological disparity pattern along the Neanderthal clade, one might consider that the later genetic contribution of modern humans (around 100ka) should have had a similar effect on their morphological disparity (that is, they should have shown a similarly high morphological variation). Indeed, even if the introgression of *Homo sapiens* genes was lower at 100 ka (~0.5%; Li et al., 2024) than in earlier events (up to 5%), we should see a similar signal in late Neanderthals, particularly because we sampled a much more diverse collection of Neanderthals (both geographically and chronologically). This is, however, not the case and we instead observe the opposite. Hence, we left our interpretation for the drop in morphological variation between Krapina and late Neanderthals unchanged.

2. Evolutionary Theory, Variation and Disparity: Associated to clarifying the phylogenetic context of the study, a brief introduction to the concept of disparity is necessary at the outset in order to justify such analysis. Aside from the fact that the concept might be unfamiliar to a general readership, it is often confounded with the concept of variation and diversity. There is a rich theoretical history to explain patterns of disparity, primarily along ecological and developmental hypotheses and it is unclear how this study is approaching it (i.e. how it is justifying the analysis). After providing the concept's background in the Introduction, the Discussion should then refer back to that underlying theory in the context of the results. Note that for some of the readership, an analysis of morphological disparity would typically entail quantification of distance between species. In this case, the calculation is between hypothetical paleo-demes (and a recent modern human group with one fossil), so the study design needs to be properly justified in this context (see also point 3, below, on sampling strategy).

We would like to thank the Reviewer for bringing our attention to a point in our analysis setup that might not have entirely been clear to the readership and that required further attention. We have thus included a new paragraph at the end of the introduction where we define the principles of disparity analysis and how these would help us in testing the likelihood of the four evolutionary scenarios for the Neanderthal clade. The amount of text devoted to these explanations is partly limited by the need of providing an adequate phylogenetic and evolutionary background in the Introduction, while considering the word limit for *Nature Communications* of 5000 words for the main text (excluding Materials and Methods, Acknowledgements, Abstract, and References). For this reason and to better guide the readership throughout our manuscript, we further expanded the *Disparity analysis* section in the Materials and Methods. Here we now thoroughly explain the reasoning behind the setting of our analysis, how we investigate within deme morphological diversity and compare it to those of the other demes included in the study, and the justification for our disparity metric selection (i.e., distance to group centroid, variance, and pair-wise distance). Lastly, we included additional interpretation about disparity patterns in the discussion, delving more in depth into the possible causes and consequences of the observed patterns. We believe that this will guide more adequately the reader throughout our reasoning and clarify better which is the background upon which we based our interpretations.

3. Sampling: while access to data might preclude the inclusion of other fossils, some discussion is necessary on how/whether doing so could change results. What other fossils could be included? What are the considerations of the geographical and temporal sampling (i.e. what spatiotemporal “gaps” need to be filled?)?

Following the Reviewer's comment, we were more assertive in the discussion about which are the gaps that still need to be investigated in the future and how it would be possible to do so. Although we were already mentioning that the time range between MIS 11 and MIS 5 is critical to achieve a better

understanding of morphological disparity patterns in the Middle Pleistocene, we added reasoning about how it might be possible to tackle this question. Similarly, we currently do not have an appreciation of the evolutionary history of Neanderthals in Asia. Indeed, besides Kebara and Amud we did not have access to any additional Asian specimen. We have thus included a sentence in the discussion to highlight that their inspection might be of great interest for furthering the analysis Neanderthal evolutionary history.

For the analysis on disparity and the inferences drawn therein, what are the implications of not sampling a hypothetical ancestor of the Neanderthal and sapiens lineage (i.e. heidelbergensis/rhodesiensis/bodoensis and/or ergaster/erectus)?

In the first place, we would like to point out that our analysis of morphological disparity is not equal to a disparity through time, as the Reviewer already pointed out in a few previous comments. Hence, we do not estimate the amount of disparity at the ancestral nodes of a phylogeny, but rather sample it only in those ranges of time for which we have fossil evidence. Hence, we consider that the lack of earlier specimens is not hindering the soundness of our interpretations, especially those proposed for the latest stages of the Neanderthal clade evolution.

However, we agree with the Reviewer that for making any reliable inference about the amount of morphological disparity that characterized the species ancestral to the Neanderthal-modern human-clade, we would need to inspect several additional specimens belonging to *Homo heidelbergensis/rhodesiensis/bodoensis* and/or to *Homo ergaster/erectus*. However, this goes way beyond the scope of the present analysis, that is: inspecting changes in morphological variation occurred within the Neanderthal clade. The reasons for limiting our analysis to the Middle and Late Pleistocene Eurasian are twofold. First, there is no current consensus about the phylogenetic position of several of the individuals that we might be able to sample (e.g., Kabwe) and thus it would imply a large degree of uncertainty at the time of interpreting the results for these specimens. Until the phylogenetic affinities of these specimens are further inspected and clarified, we prefer to maintain a conservative position and only include individuals and assemblages with more established affinities. Second, we lack access to key specimens both in early and later stages of the Neanderthal-and-modern-human-clade. This is because they do not preserve a bony labyrinth (e.g., Bodo and the *Homo antecessor* specimens from Gran Dolina, Atapuerca), they are currently analyzed by other researchers (e.g., Jebel Irhoud, Omo I and II and several others, which are key for understanding the evolution of *Homo sapiens*), or due to issues in performing computed tomography analyses (e.g., the cranium from Buia, Eritrea).

For the reasons above, we discarded the idea of inspecting the patterns of morphological variation for the whole Neanderthal-and-modern-human clade, at least at the present time. Conversely, we decided to rely only on the population size estimates available from genetic data (e.g., those presented in Rogers et al. 2017), which constitute the starting point for our reasoning as we explain in the Introduction of the manuscript.

Similarly, how might the inclusion of early Middle-Late Pleistocene anatomically modern humans and later Late Pleistocene modern humans inform analyses on disparity and variance (i.e. arguably two separate demes)?

We concur with the Reviewer's opinion that it would be of utmost interest to understand how Middle-Late Pleistocene anatomically modern humans and later Late Pleistocene humans would influence our analysis. Unfortunately, we have limited access to these specimens (since they are currently under study, as specified in the response above) and thus were not able to include many of them in the present study. We nonetheless tried to increase the amount of fossil modern humans in the analysis by adding another Upper Paleolithic individual from the French site of Laugerie-Basse. Notably, the inclusion of this specimen did not substantially modify the patterns in shape diversity observed in modern humans.

What is the effect of only including one fossil in the recent modern human group? Is this justified because the entire Cro-Magnon fossil assemblage is assumed to be recent (see Fu et al. 2013 for radiocarbon date, <https://doi.org/10.1016/j.cub.2013.02.044>), because the date range between specimens is relatively small compared to the possible range in the other groups, or for another reason?

The reasoning behind the inclusion of Cro-Magnon 1 was that of trying to increase the amount of time covered by our sample and make it better comparable to those of Sima de los Huesos, Krapina, and, particularly, late Neanderthals (the latter widely ranging geochronologically). As mentioned above, we unfortunately do not have access to many other Upper Paleolithic individuals, but we did include a new specimen from Laugerie-Basse to increase somewhat more the inspected phenetic pool for modern humans. The composition of the modern human sample notwithstanding, we would like to point out that *Homo sapiens* individuals only constitute the comparative sample in the present analysis and are used here as a term of reference for a population that was recently affected by a genetic bottleneck event.

4. Methodology: because the diffeomorphometry (DMorph) method relies on the comparison of surfaces, a calculation of measurement error must be presented for (a) the segmentation procedure used to develop the surfaces and (b) for differences between micro-CT and medical CT data. The former (a) is straightforward and would consist of segmenting a bony labyrinth multiple (3+) times to assess any surface variation within the same individual, noting also variation in thresholding parameters or other procedures (e.g. manual delineation of structures). The latter (b) would consist of segmenting the bony labyrinth for an individual where both micro-CT and medical CT data is available and assessing surface variation between the two. While both error tests are important, the latter is of particular concern given that medical CT data comprises voxel sizes that are non-isometric. For relatively large anatomical elements (e.g. a cranium, where a relatively high number of voxels are represented), measurement error will likely be negligible; however, for relatively small anatomical elements (e.g. the bony labyrinth, where a relatively low number of voxels are represented), measurement error might be significant. Non-isometric medical CT scans will inevitably create a bias along the Cartesian direction comprising the larger voxel size, with variation added if the specimens are scanned in different directions. Therefore, it is important to show whether the combination of micro-CT and medical CT data affects how the surface is quantified when applying DMorph. This should be tested in one individual at minimum but ideally in multiple individuals in order to show that variation associated with voxel type is minimal and relatively less than the variation otherwise observed per hominin group sampled. In other words, the variation associated with data type (micro-CT vs medical CT) should be less than the variation from any of the hominin groups. Note also that the degree of error can have an impact on the visualization of deformations along PCs. For example, surfaces based on non-isometric voxels might appear "...more voluminous and inflated..." (line 178) than surfaces based on isometric voxels.

We thank the Reviewer for the valuable comments and for the precise directions to overcome the concerns he raises, some of which coinciding with those expressed by Reviewer #2. To address them, we performed two separate tests. With the first test, we assessed the segmentation-related errors in our analysis by segmenting three times four individuals (one modern extant human [F01], one Krapina Neanderthal [Krapina 39.3], one Late Pleistocene Neanderthal [Amud 7], and a Sima de los Huesos human [Cr.3]). The specimens were selected due to their different state of conservation (Amud 7, Cr.3, and Krapina 39.3 showing abundant sediment in the bony labyrinth cavities, whereas F01's inner ear was free of any inclusion) and/or different scanning parameters (F01 and Amud 7 being microCTs, whereas Cr.3 and Krapina 39.3 were scanned using a medical CT scan). The segmentation repetitions were performed by the same author (A.U.) on separate days with an interval of a week between each attempt to ensure the independence of the segmentation attempts. Each attempt relied on the method described by Ullrich et al. (1980) to assess the gray value threshold to be used in the segmentation (reference to

this method was added in the revised Materials and Methods section). This method, known as half-maximum height (HMH) method and frequently used in the literature (e.g., Spoor and Zonneveld, 1998; Kirk and Gosselin-Ildari, 2009; Conde-Valverde et al., 2019), considers the highest and lowest gray values on either side of an interface bony canal boundary and then uses the average as threshold value. In our case, the gray values were assessed on the compact bone surrounding the slender portion of the anterior canal lumen using the Plot Profile function of ImageJ (Schneider et al., 2012). Please note that in all cases except the modern human F01, additional manual corrections were required due to the presence of sediments in the semicircular canals and vestibule of the fossil specimens. The repetitions were then analyzed using our diffeomorphometric method. The measurement error was assessed by performing a PCA of the shape data of the repeated individuals (currently show in Supplementary Figure S4). The results demonstrate that regardless of the CT type (CT or microCT) and sediment presence/absence, the three repetitions plotted extremely close to one another. From an analytical perspective, we computed the Euclidean distance between the repetitions in the first three PCs (98.6% of variance) and compared the average distance of each of the repeated individuals and the mean interindividual distance. Results show that in each case the inter-repetition distance was much smaller than interindividual distance ($\leq 2\%$), thus implying that shape variation caused by segmentation processes can be regarded as negligible.

The second test aimed to assess the influence of CT data type on group disparity, being this a main concern also for Reviewer #2. To do so, we performed an analysis of the full sample (including the newly added individual of fossil *Homo sapiens*) in which we substituted the microCT data of five of the modern human individuals with those deriving from a standard CT scan. This was only possible for the individuals derived from Wimmer et al. (2019) as for all other specimens we possessed only a type of CT data (being it microCT or standard CT). However, we artificially increased the voxel size (decreased the resolution) of one Late Neanderthal individual (La Ferrassie 2) by resampling the original scan in Fiji (ImageJ) to a voxel size of 0.156x0.156x0.4mm. In this way, we increased the voxel size range for this group (the least variable) so that it would become similar to that of Krapina (the most variable). We then performed morphological disparity analyses on this modified sample (which also included the surfaces of Cr.6 and Kr38.20 derived from the resampled scans; see our response above to another comment of Reviewer #2) and compared the results with those obtained on the original data. The results showed an identical distribution of morphological disparity in the modified sample analysis (Supplementary Figure S3), thus demonstrating that CT data type and voxel size do not affect our results.

Minor points:

5. Methods: clarify the mesh [pre]alignment process (lines 384-390). While unclear, I understood this to mean that all sampled surfaces are superimposed manually rather than in a standardized fashion that relies on, for example, homologous landmarks. It is not encouraging to read that the meshes were "...roughly prealigned..." so an error test would also be in order here if the pre-alignment process is not standardized.

We thank the Reviewer for highlighting a point where our methodology was not totally clear. The pre-alignment process we mentioned is only required to avoid issues with differences in scale and/or initial orientation. Regardless, we also performed an error assessment in a similar way to that used to inspect the effect of segmentation replicates. Hence, we performed the automatic alignment of the surfaces using the "Align Surface" module of Avizo, with the 'rigid + uniform' option enabled starting by different initial positions and repeated the process five times for four individuals, one from each of the considered groups (F01, Amud 7, Cr.3, and Krapina 39.3). The repetitions were then analyzed using our diffeomorphometric method. The measurement error was assessed by performing a PCA of the shape data of the repeated individuals (currently show in Supplementary Figure S4). Our result shows a very close packing of the individuals when the shape data is analyzed by means of PCA. As for the segmentation, the error was

assessed by inspecting the Euclidean distance between the repetitions in the first three PCs (99.9% of variance). The average distance between the repeated individuals was always much lower than the mean interindividual distance ($< 1\%$), allowing us to discard a possible influence of the initial position of the surfaces in the alignment process.

6. Methods: justify the PC selection process. Was choosing PCs explaining ~90% of variance an arbitrary cut-off, those explaining $>1\%$ variance, or something else? See, e.g. Reyes-Centeno, Harvati, and Jäger 2016 (and references therein) for example of systematic method for PC selection (<https://doi.org/10.1038/srep36645>).

Following the Reviewer's suggestion, we were more clear-cut about the decision made to select the PCs to be included in the disparity analysis. More in detail, in the current version of the manuscript we selected only those PCs explaining $>1\%$ of variance. The reason behind this decision is justified by the results obtained in the paper suggested by the Reviewer, where the authors highlighted how their bootstrap-component-based analysis conformed to "the common practice of excluding PCs that explain less than 1% of variance, as these components explain less of the variance than the original shape variables". Hence, we included the first 19 PCs, accounting for a total of 86.7% of total variance.

7. Results, PCA: when discussing the results, (a) be specific about how the overlap of groups is determined (i.e. based on the convex hull) and avoid using "discriminate" since this is not a linear discriminate or canonical variate analysis here. Put another way, what would the results look like if you visualize 95% confidence intervals instead of convex hulls? It is fine to visualize with only convex hulls, but then the description and interpretation of results should be specific to these analyses. In addition, (b) quantify how clusters can be compared. For example, I don't see the "tight(er) cluster[ing]" (lines 214, 221) that is interpreted for the RMH and LN groups. I might be convinced if the area (in the 2D case of 2 PCs) or volume (in the 3D case of 3PCs) of the convex hulls is quantified and compared. Otherwise, it would suffice to refer to the group dispersion results (but see below).

Following the Reviewer's comments, we decided to more specific about the range occupied by each group along the PCs visualized in Figure 3 (PC1-4) and Supplementary File 1 (PC1-3). More in detail, we included 95% ellipses for the 2D bivariate plots shown in Figure 3 and 3D ellipsoids in Supplementary File 1 (note that ellipsoids and 3D convex hulls embedded in the html file can be turned on/off to freely inspect the 3D morphospace, if needed).

8. Results, group dispersion: specify how significance in group dispersion values between two groups was determined, i.e. what statistical test was used to quantify significance.

As specified in the response to one of the Reviewer's #2 comments, we have decided to replace the group dispersion analysis performed on the raw shape data with a pair-wise Euclidean distance analysis based on PC1-19. This is due to drawbacks related to the direct analysis of shape data for assessing morphological diversity (please see the response above for additional information).

9. Line 111: rather than highlighting the bony labyrinth as a "...marker of population dispersal in modern humans," it would be better in this context to justify its use as "...population history of modern humans" to highlight its use in intra- specific variation.

Following the Reviewer's comment, we modified the wording of the sentence. We thus added "history" but did not remove "dispersal". This is due to the (paleo)biogeographical implications of word dispersal, which are also key for the present paper and was included in the title of the cited reference (i.e., "Human bony labyrinth is an indicator of population history and dispersal from Africa") Hence, the sentence now goes as "as a marker of population history and dispersal in modern humans".

10. Lines 118-119: amend to "...sagittal labyrinthine index, which quantifies how the lateral anal plane intersects the posterior canal superiorly (Fig. 3d-f)" (referring to the Figure will be useful for non-specialists, see also comment 30, below, on figure labels).

We respectfully disagree with the Reviewer on this point. Here rather than explaining what the labyrinthine index is, we are reporting a derived condition for the Neanderthals (the arrangement of the lateral canal relative to posterior and anterior canals). This is why we included an explanation of the derived Neanderthal condition right after the i.e. Similarly, the Reviewer suggests adding a reference to our Fig. 3. While this would be logical in the Results and Discussion, the shapes shown in Fig. 3 are mean shapes obtained from our DMorph analysis. Hence, in that way we would anticipate part of the analysis in the introduction, which is something that we would like to avoid. However, we agree that our paper has to be easily accessible to a readership with a broad interest, such as that of *Nature Communications*, and thus we included a new panel (with lettering b-e) where we show a selection of semicircular canal and vestibule 3D models, with one for each of the considered groups. When also included in the picture the labels for the semicircular canals, common crus, and vestibule. We consider that this will help the readership to understand better our analysis and also address another concern (comment 30) by the Reviewer.

11. Line 121: change "...is not present in..." to "...is not present to the same degree..." (or similar amendment)

Although we agree with the Reviewer that SLI is a continuous character and thus it is not easy to clearly define a limit for when it should be considered Neanderthal-like or not, we would like to bring to the Reviewer's attention the work by Quam et al. (2016), where this feature was carefully measured and assessed. Here, four individuals out of 14 do fall within the Neanderthal range (close to the lowest limit of the range, something that occurs also with several modern human individuals) whereas all other individuals show a lower position of the lateral semicircular canal relative to the posterior one (i.e., they have a SLI < 53). Notably, the average SLI of SH humans (49.2 ± 7.1) is significantly lower than that of later occurring Neanderthals (63.4 ± 5.7 , including Krapina) at $p < 0.001$, whereas it does not significantly differ from that of modern humans (50.6 ± 5.4). The authors concluded that SH humans are plesiomorphic in their SLI relative to the derived condition of later occurring paleodemes of the Neanderthal clade (Krapina included and possibly as early as Reilingen; Spoor et al., 2003). We thus follow this interpretation (which was not challenged by more recent contributions) and did not change the wording for line 151 (former line 121).

12. Lines 123-124: change "...Aroeira 3...SH humans and is also from the Iberian Peninsula," to "...Aroeira 3 from Portugal...SH humans from Spain are geographically proximate within the Iberian Peninsula," (or similar amendment) to provide reference to non-specialists

In view of the Reviewer's suggestion, we have rephrased this period (and part of the previous one) to make sure that the readership is given the greatest amount of information possible. Hence, we added that the Sima de los Huesos humans were recovered from a Spanish site (line 152 of the track-changes version); we included that Aroeira 3 represent a partial calvaria from Portugal (lines 153-154 of the track-changes version); this period was also separated in two to better convey and remark in which features this individual appears more primitive than its penecontemporaneous ones.

13. Line 125: remove "much" or otherwise quantify descriptor

The word "much" was removed from the updated version of the sentence.

14. Line 126: change “displaced” to “placed” (the former might connote a pathology or functional trait that reduces fitness)

The word was changed following the Reviewer's suggestion.

15. Line 126: change “instead” to “thus” to provide continuity with previous sentence. Note also that some readers might otherwise mischaracterize the use of “primitive” (see also point 1, above, on phylogenetics)

The change was implemented according to the Reviewer's comment.

16. Line 367: change “...430 ka” to “...~430 ka”

We included the tilde here and other instances.

17. Line 133: change “...this study...” to “...that study...” (to clarify the preceding reference rather than the present original study)

We followed the Reviewer's suggestion and included the edit.

18. Lines 144-147: amend incomplete sentence (e.g. “Two geochronologically constrained populations...were sampled to maximize...”)

We amended the wording which now goes as: “Two relatively large and chronologically constrained samples—a Middle Pleistocene population from SH and a late Middle Pleistocene population from Krapina—and a spatially and chronologically heterogeneous sample of late Neanderthal individuals. The latter was designed to maximize the morphological variation for Late Pleistocene Neanderthals.”

19. Lines 154, 155, 166: amend change equal symbol to approximately equal symbol (e.g. line 154: “SH = Kr = LN” to “SH $\approx$ Kr $\approx$ LN”)

We agree with the Reviewer. The requested changes were implemented.

20. Line 208: consider changing “...hiatus between RMH and the Neanderthal clade...” to “...space between the convex hulls formed by RMHs and Neanderthals....” Note that the term “hiatus” is awkward in this context and might be interpreted to imply an associated hiatus in the archaeological record.

The word “hiatus” was replaced by “separation” in the updated version of the manuscript.

21. Lines 218-220: disambiguate this sentence by removing the parenthetical information and including it so that the comparisons are properly understood.

The sentence was revised, so that the names of the Upper Paleolithic individuals are now placed outside the parentheses.

22. Lines 380-381 and Table S1: Specify in Table S1 whether the right side of the bony labyrinth was used and make reference to it in Lines 380-381. In Lines 380-381, quantify the percentage of cases where mirroring was applied.

We have now included a new column to specify the laterality of the specimens, as well as the % of mirrored specimens in the main text.

23. Line 414: clarify what the “DeterministicAtlas option” refers to: is this the sample-average template that enables the calculation of momenta deformation vectors?

We have changed the wording according to the Reviewer's suggestion and clarified that: ‘*estimate*’ is a command for the Deformetrica v.4.3 software; “DeterministicAtlas” is the diffeomorphic function to

mathematically model the deformations between each specimen and the template; and then generate an *atlas of momenta* (as explained in lines 552-556).

24. Line 416: clarify what the raw shape data is (vector lengths?)

As asked by the Reviewer, we clarified the PCA was performed based on the atlas of momenta (or deformation fields) obtained from the DMorph analysis. To further clarify the operative processes, we are now providing the R code that was used to perform all the analyses.

25. Line 425: briefly explain how disparity is calculated (i.e. how PC scores are treated).

Following the comments of all the Reviewers, we have included additional explanations regarding the calculation of disparity and of the metrics we employed to understand the distribution of individuals within the multidimensional shape space.

26. Line 678: delete “Resume of the...” (or change to “Résumé of the...”/“Summary of the...”)

The caption for figure 2 now directly starts as “Scenarios of phenetic variation...”

27. Line 685: delete “Three-dimensional...”

We are grateful to the Reviewer for catching this! Views of our three-dimensional PC plot were included in a previous draft of the main text of the manuscript but are currently not necessary. We also removed a duplicate of the word “analysis” in the same caption.

28. Lines 690-691: change “fossil modern human” to “recent modern human” (see also point 3, above, on sampling strategy)

The caption of Figure 3 was changed to clarify that orange corresponds to modern humans (both recent and fossil [Cro-Magnon 1 and Laugerie-Basse]). We thus removed the word “fossil” and left “modern humans” only. Regarding the sampling strategy, we provided above a thorough response to the concerns of the Reviewer.

29. Table 2: add label descriptor for Cro-Magnon 1

We thank the Reviewer for noticing this lack of information. The identifier of Cro-Magnon 1 (as well as that of the newly included Laugerie-Basse) is now included in Table 3.

30. Figure 3: add labels to one of the canal figures (e.g. Fig. 3c) which conveys the names of the canals. This will aid a non-specialist reader in understanding the morphology.

We strongly agree with the Reviewer that the readership might not be too familiar with the morphology of the vestibular system. Thanks for the suggestion! Hence, we decided to add information about the position and naming of the structures composing it in a figure. However, as explained in one of the points above, we have done so in Figure 1 rather than in Figure 3, as the latter was designed to give the results of the analysis rather than introducing it.

31. Figure 3 Legend, Lines 685-691: clarify “HA” label (Cro-Magnon)

We agree with the Reviewer that it was not totally clear in which part of the morphospace the fossil modern human specimen(s) fell in our previous version of Figure 3. Hence, we have revised the design of the figure extensively and added the specimen names for these individuals as well. We did not specify the complete ID of these specimens as it would be too lengthy to correctly fit into the overall design of the figure. Nonetheless, we included this information with the complete MNHN label in Supplementary Table 2.

Hugo Reyes-Centeno

Assistant Professor, University of Kentucky

Department of Anthropology

211 Lafferty Hall, Lexington, KY 40506 USA

William S. Webb Museum of Anthropology

[1020 Export St, Lexington, KY 40504 USA](#)

heva.as.uky.edu | educelab.engr.uky.edu

Additional refs.

Arsuaga, J.L., *et al.*, 2015. Postcranial morphology of the middle Pleistocene humans from Sima de los Huesos, Spain. *PNAS* **112**, 11524–11529.

Braga, J., *et al.*, 2019. Cochlear shape reveals that the human organ of hearing is sex-typed from birth. *Sci Rep* **9**, 10889.

David, R., Bronzati, M., Benson, R.J., 2022. Comment on “The early origin of a birdlike inner ear and the evolution of dinosaurian movement and vocalization”. *Science* **376**, eabl6710.

Harvati, K., Weaver, T.D., 2006. Human cranial anatomy and the differential preservation of population history and climate signatures. *Anat Rec* **288A**, 1225–1233.

Kirk, E.C., Gosselin-Ildari, A.D., 2009. Cochlear labyrinth volume and hearing abilities in primates. *Anat Rec* **292**, 765e776.

Manica, A., Amos, W., Balloux, F., Hanihara, T., 2007. The effect of ancient population bottlenecks on human phenotypic variation. *Nature* **448**, 346–349.

Mennecart, B., *et al.*, 2023. Ruminant inner ear shape records 35 million years of neutral evolution. *Nat Commun* **13**, 7222.

Osipov, B., *et al.*, 2013. Sexual dimorphism of the bony labyrinth: A new age-independent method. *Am J Phys Anthropol* **151**, 290–301.

Ponce de León, M.S., *et al.*, 2018. Human bony labyrinth is an indicator of population history and dispersal from Africa. *PNAS* **115**, 4128–4133.

Roseman, C.C., Weaver, T.D., 2007. Molecules versus morphology? Not for the human cranium. *BioEssays* **29.12**, 1185–1188.

Reyes-Centeno, H., *et al.*, 2014. Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *PNAS* **111**, 7248–7253

Schneider, C.A., Rasband, W.S., Eliceiri, K.W., 2012. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671e675.

Spoor, F., Wood, B., Zonneveld, F., 1994. Implications of early hominid labyrinthine morphology for evolution of human bipedal locomotion. *Nature* **369**, 645–648.

Stoessel, A., *et al.*, 2024. Spirals through time: evolutionary history of the modern human bony labyrinth. *9th ESHE Meeting Abstract Book* 168.

Uhl, A., Karakostis, F.A., Wahl, J., Harvati, K., 2020. A cross-population study of sexual dimorphism in the bony labyrinth. *Archeol Anthropol Sci* **12**, 132.

Urciuoli, A., *et al.*, 2020. The evolution of the vestibular apparatus in apes and humans. *eLife* **9**, e51261.

Ward, D.L., *et al.*, 2020. Can bony labyrinth dimensions predict biological sex in archaeological samples? *J Archeol Sci Rep* **31**, 102354.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author)

The new version of the manuscript, "Semicircular canals shed light on bottleneck events in the evolution of the Neanderthal clade" addresses most of my and the other reviewer's comments appropriately. Moreover, it stands out as a much-improved version of the manuscript. I would like to thank the authors for their great work. I am convinced that this is an excellent contribution to Nature Communications, and I would like to congratulate the authors for their great job. However, I still have some minor concerns, which I believe won't be difficult for the authors to reconsider and resolve at this stage. I hope these points help strengthen some aspects of the manuscript that still need attention. All my comments refer to the PDF version with tracked changes.

We are very grateful to Dr. Menéndez for the positive remarks about the revised version of our manuscript and we acknowledge her for the constructive comments. Thanks!

1. I agree with the authors that the morphological variation of the bony labyrinth is an excellent proxy for discussing random factors in evolution and, as such, random genetic variation for reconstructing population or evolutionary history. I appreciate that the authors have improved many parts of the text following my previous comments on this. However, the way this has been phrased in some passages of the manuscript is still ambiguous. In general terms, whenever they mention "genetic bottleneck," I would be more cautious and replace it with "bottleneck" (please check throughout the manuscript). This is mainly because not all genetic variation is random. This is perfectly expressed, for example, in lines 223-225 of the manuscript. Therefore, I suggest changing the wording, particularly in these cases: 1) In the abstract, line 35, instead of "genetic variation," it should say something like random or neutral factors; 2) In the abstract, line 38, instead of "genetic variation," it should say something like evolutionary history/population history; 3) In the Introduction, line 188, instead of "genetic variation," change to random or neutral variation.

We agree with the reviewer about this point and carefully checked the manuscript to change the wording as suggested. More in detail:

L35 and L38 – The wording was changed to "The goal of this study was to quantify the degree of disparity in the Neanderthal clade. We hypothesized that a reduction in bony labyrinth disparity is indicative of the underlying genetic variation resulting from bottleneck events", also following the suggestions of Reviewer #3.

L188 – We removed the word "genetic" from "genetic bottleneck"

We did not change the instance at L122, L326, L372, L379-380 (track changes version of the revised manuscript) as these were not statements of ours but rather a reference to previously published results.

2. In relation to chronology, I think it would be clearer for the readers, especially considering the wide audience of Nature Communications, to include, in brackets, the equivalent chronology to the MIS stage whenever possible as they are mentioned throughout the text, particularly in the Introduction and Discussion. I know that this information is available in Figure 1, but personally, I think it will make the narrative clearer

Following the reviewer's suggestion, we included the chronology for the marine isotope stages the first time we mentioned in the text.

3. Introduction, lines 130-131: The authors argue that "periods of drastic decrease in disparity might represent strong evidence for the occurrence of genetic bottlenecks." The first question that comes to my mind is: A decrease in disparity means an increase in affinities among the groups analyzed, so why couldn't this be the result of other processes, for example, gene flow or shared/parallel adaptations?

Based on the reviewer's feedback, we realized that our previous phrasing lacked clarity. First, we would like to point out that in the present analysis we considered the within-group morphological disparity, rather than across groups. Since the groups are arranged chronologically, a difference in disparity between consecutive groups indicates a change in the amount of morphological diversity that occurred somewhere during the time interval separating those groups. To make this point clearer, we explicitly stated (both in the Materials and Methods and in the Introduction) that disparity in our analysis always refers to the amount of morphological diversity observed within each group (i.e., within-population).

Notwithstanding the above, we now explicitly consider the possibility that other evolutionary processes might play a role in increasing or decreasing the amount of morphological diversity.

4. Discussion, lines 401-403: The authors argue that "the increased phenotypic diversity of Krapina...might be the result of neutral phenotypic evolution..." It is not clear to me why the authors suggest neutral evolution in this context, which is usually the result of drift or mutations. Usually, in small and isolated populations, drift is associated with a decrease in genetic variation, while mutations happen at such a slow rate that they would not be sufficient to explain such changes in diversity. I would instead suggest other factors to explain this, like the ones suggested in the following sentences (lines 404-416: natural selection, gene flow, phenotypic plasticity). Unless I misunderstood and when they refer to neutral evolution meaning drift, they imply the fixation of unique

traits within a population, increasing the diversity between the compared groups. If this is the case, this clarification should be added to the manuscript.

The reviewer is right as we wanted to refer to drift as a process of neutral evolution. The revised version now directly mentions drift to avoid misconceptions. Thanks for catching this!

5. Discussion, line 482: The authors mention “discarded variation,” which I think sounds a bit odd. I would replace it with retained/relict/remnant variation, if I understand correctly what they are trying to convey here.

Following the reviewer’s suggestion, we changed “discarded” for “relict”. Thanks!

Minor comments:

Line 61: change to “sometimes regarded as a separate species..”, changing the position of the “a”

Done.

Line 495: change to “the evidence provided by semicircular canal shape variation”, deleting the “and”

Done.

Line 499: I think the authors mean “more rapidly/fast” instead of “more frequently”

Done.

Lumila Paula Menéndez, University of Vienna

Reviewer #2 (Remarks to the Author):

The authors have gone to great lengths to address my and the other reviewers’ comments and concerns on the first draft of the manuscript (NCOMMS-24-28932). I agree with and am convinced by all the responses and decisions relating to the comments raised by myself and the other reviewers.

I greatly appreciate the in-depth analysis and responses to my original questions about scan resolution as well (and apologies that I had not already seen your previous work that addressed this issue).

I have no major concerns about this new draft, and only have a few minor comments or suggestions which are detailed below.

Altogether, I thought the original manuscript was quite strong and the revised manuscript is even stronger, so I will be happy to see this paper in print soon. I have learned a lot about the vestibular system (as well as later Pleistocene human evolution and analytical methods) from this paper, thank you!

We are most grateful to the reviewer for the very positive remarks and consider that the constructive suggestions and remarks by the reviewer considerably enhanced the robustness of our paper, thanks! We have addressed all the minor comments as specified in the point-by-point response below.

Minor comments

Abstract, line 35 of revised text: change “drop” to “reduction”

We have changed “drops” for “reductions”.

Line 81 of revised text: change “leading” to “led”

The sentence was changed to accommodate the reviewer’s suggestion and now goes as “...which ultimately led to the...”

Line 137: add “s” to other ◊ “Among others...”

Done.

Line 137: “clockwise” should be replaced or specified as it will have a different meaning for the right vs. left sides. E.g. Spoor et al. (2003, p. 158) specified “When viewing the lateral aspect of the left temporal bone and labyrinth this morphology can be seen as a joint clockwise rotation....”

We agree with the reviewer and add “... (when the left bony labyrinth is observed in lateral view) ...” to disambiguate our statement. Thanks!

Along these lines, perhaps Figure 1 should specify that images depict the left vestibular system.

Thanks for catching this! This information was indeed required and inadvertently omitted in the previous version.

Line 245: the word “modern” is missing the ‘n’

Fixed.

Line 299: there is a large space between “customarily ... considered”

We have removed the tab that was mistakenly placed there.

Line 307: The sentence “Our results further remark on the post-Sima...” I think it would be clearer if it was stated or reviewed exactly what the results tell. I.e., “Our results show that a late Neandertal-like morphology evolved...” highlighting how this anatomy appeared/changed between Sima, Krapina, and the final late Neandertal sample. (this explanation does appear later in lines 395–397)

To better convey the evolutionary scenario that our results suggest, we included a short statement at the end of the sentence to specify that Krapina already shows a great frequency of a late Neanderthal-like lateral/posterior canal arrangement.

Line 316: Change “conception of” to “argument for.” Note also that “drastic” appears in two sentences in lines 316 and 317, so stylistically one of these instances could be changed to similar word like “extreme” or “severe.”

The change was implemented in accordance with the reviewer’s suggestions. We also modified the second “drastic” for “...the Neanderthal lineage underwent a severe population reduction...”

Line 340: Change “Tecent” to “Recent”

We thank the reviewer for catching this typo. Fixed.

Line 364: remove the comma (,) between “variation also eliminates”

Fixed.

Reviewer #3 (Remarks to the Author):

The revisions have greatly improved this manuscript and properly addressed comments and suggestions from the earlier draft. Some minor points of revision are enumerated below.

We sincerely appreciate the reviewer’s positive feedback and recognition that the revised version of our manuscript effectively addresses their concerns.

1. Methods, sexing: Briefly explain the biological sexing of the recent human samples (e.g. karyotyping, assigned-at-birth, self-reported, etc.)

A sentence was added to refer the reader to Supplementary Table S1 and to specify that sex, when available, was established at birth and reported in the metadata of the analyzed collection.

and cite the source for the fossil samples (e.g. with superscript references in in Table S1), noting limitations.

The references for the sex of the fossil sample were included in the revised version of the manuscript. We also noted the current limitations by adding “although information is not available for several individuals” in the main text when citing Supplementary Fig. 2

If any of the fossils were assigned a sex for this study, explain the criteria used and its limitations.

None of the fossils was assigned sex during the current study and sex information was either extrapolated from the literature or from the metadata associated to the extant sample.

Consider adopting the term “sexual differences” rather than “sexual dimorphism” for this general science journal (see, e.g., review by Khramtsova et al. 2019. doi:10.1038/s41576-018-0083-1).

We thank the reviewer for highlighting this point. The requested changes were implemented.

2. Within the text, clarify the PC selection process. This was explained in the rebuttal, but not explained in the main text.

A sentence explaining the selection process and a reference were added.

3. Abstract, lines 35-38: this sentence has a tone of speculation, so consider rephrasing this sentence for clarity and making an explicit hypothesis instead, e.g. “The goal of this study was to quantify the degree of phenotypic disparity in the Neanderthal clade. We hypothesized that a reduction in bony labyrinth disparity is indicative of the underlying genetic variation resulting from bottleneck events.”

We thank the reviewer for the suggestion which was implemented in the revised manuscript.

4. Abstract, line 42: change “...providing strong evidence for...” to “...supporting our hypothesis...”

The change was implemented.

5. Citation edits: superscript citations should not be accompanied by additional text. If additional text is necessary, then include it parenthetically as a normal script and the citation number preceded by “Ref.” For example, line 66 would be: “...(see Ref. 24 and references therein for a review). In most/all instances, the accompanying text can be excluded and the citations can simply be in superscript.

The format of the superscript citations was modified accordingly.

6. Introduction, line 80: change to “...late Neanderthal morphology...” or “...late Neanderthals’ morphology...”

Thanks for catching this! We removed the 's' of Neanderthals.

7. Introduction, line 82: remove reference to *H. heidelbergensis* since previous paragraph has established the possibility of *H. antecessor* as ancestral

Although the reviewer is right regarding our statement regarding the possibility that *Homo antecessor* might be ancestral to the Neanderthal clade, we would like to remark that currently no remains pertaining to this species have been identified in the Middle Pleistocene. We thus preferred to maintain the reference to *H. heidelbergensis* here. We also consider that this is warranted by the explanations we provide in the text regarding how this species might be conceived and by the fact that we only refer to European Middle Pleistocene populations.

8. Introduction, line 146: change "other" to "others"

The change has been implemented.

9. Introduction, lines 164-167: amend for clarity, e.g. "...bony labyrinths highlighted...variation in the latter group."

We amended the text accordingly.

10. Throughout the manuscript, use site names of Sima de los Huesos and Krapina as an adjective for the sample/group/population rather than a noun, e.g.:

a. Introduction, lines 191-194, 198-202, 203-207: amend for clarity, e.g. "...after the Sima de los Huesos population..."

b. Results, lines 238-240: amend for clarity, e.g. "...in the Sima de los Huesos, Krapina, and late Neanderthal samples."

c. Results, line 256: amend for clarity, e.g. "...in the Krapina sample relative to the late Neanderthals sample."

d. Results, line 302: amend for clarity, e.g. "The Krapina sample is more variable..."

e. etc.

The use of site names was revised throughout the text.