nature portfolio

Peer Review File

Comparative genomics sheds light on mammalian and avian gene regulation and phenotypic evolution

Corresponding Author: Dr Lucia Franchini

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)
Review of Comparative genomics sheds light on mammalian and avian gene regulation and phenotypic evolution

While I found the premise and the general scope of the results pretty interesting, I found it hard to follow the methods used to generate the datasets and so have several questions as to what exactly was done. While there were code associated with the methods, they were not easily interpretable, particularly the early element detection and filtering steps (steps on Github that is not in the Rmarkdown file). I shall reserve final judgements on some of the results and discussion pending more clarification of the methodology.

My first major question is this:
Why do you expect accelerated elements to be functional or functionally relevant? My interpretation of the logic is that conserved elements are what are important since conservation pertains to potential functional relevance (as transcriptional factors/binding sites). Acceleration is just the lack of conservation, and it may or may not still have any functional relevance. What you can infer from acceleration is that something functional that was conserved across time is not conserved anymore (hence probably nonfunctional), it does not necessarily mean that it has adapted to a new function. If an element is really functionally relevant then you will may see rapid acceleration at the base of a lineage (say all birds) but then subsequent conservation (since they are functional). I think you are looking for things that are accelerated and have stayed that way across the family (unless I am mistaken). Also, conserved regions tend to move around as motifs for binding change, new regions evolve. So, loss of conservation does not necessarily mean that there has been functional change. It could easily mean that the constrained region has moved somewhere else.
With that, I will present some of the methodological questions that I had based on what you have in your methods section:
“Because there were no multiple alignments that included a large number of mammals, birds,
and other amniotes at the time of this study”
> While I can see that based on the dates on the Rmarkdown file much of the results were generated and analyzed in 2020, this statement then was probably valid. However, in the subsequent four years many new alignments have been produced. While I don’t advocate for the authors to use these new datasets, I still think some of the results from these studies must be acknowledged. In particular, the introduction should be updated with some of the studies that were derived from the Zoonomia or the B10K alignments to address convergence. I would probably include a year in that sentence above to allow readers to interpret time of study as 2020 and not 2024/5 when this paper would be published.
“Then, first, we identified conserved regions in amniotes of at least 50 bp using phastCons software”
> I can see from the code that you have generated and used a neutral rate file to use for estimation of conserved elements. However, the information of what you actually used to generate this neutral rate file is absent. What kind of marker was used for this? Or maybe you got this file from somewhere else? Please include details in the methods.
> How did you generate the conserved rates?
> Scientific names are not italicized
“A region was considered informative when at least three outgroup species, i.e., those that are not part of the set where acceleration is sought (in this case non-mammalian amniotes) were present”
> Only by looking at the Github page and looking at parts of the code was I able to figure out that you may have grouped the turtles as one (since you sampled four turtles); or maybe I am wrong? Otherwise, if you choose three outgroups how did you make sure all three were not turtles? That would definitely bias your results to turtle conserved elements. Again, I am assuming you did account for this but what you did is not mentioned in the methods.

“we defined accelerated regions in a particular lineage as those regions that exhibit acceleration signals within a functional element”
> I am confused by this, by this do you mean you are choosing accelerated elements as any element that is accelerated in any or combination of mammalian species you selected? Does it mean an element that is conserved in all species and say accelerated in primates would be classified as accelerated?
Bird acceleration
> With regards to birds, I think it is important to account for gc-biased gene conversion as well. Since, an element with high GC content can be flagged as accelerated due to gc-biased gene conversion and not due to any selective pressures.
o Liu, Anguo, et al. "GC-biased gene conversion drives accelerated evolution of ultraconserved elements in mammalian and avian genomes." Genome Research 33.10 (2023): 1673-1689.
GO analysis
> Using GO analysis without taking into account the distribution of elements will give you a skewed result. The base assumption of using the GO analysis is that conserved elements are randomly distributed throughout the genome and hence you expect a random set of background genes. However, this is not true. We know conserved elements are more common near specific genes and gene regions. And a few genes have large number of conserved elements nearbly. Here is an example, but I think there is another paper that shows this to be true in birds and mammals as well.
o Gonzalez, Paul, Quinn C. Hauck, and Andreas D. Baxevanis. "Conserved Noncoding Elements Evolve Around the Same Genes Throughout Metazoan Evolution." Genome Biology and Evolution 16.4 (2024): evae052.
So, if you take a biased set of elements, you are more likely to encounter elements that are accelerated in GO categories that are already overrepresented by conserved elements. Hence, you cannot make any inference as to whether you see a certain GO category as being enriched is due to the prevalence of large number of accelerated elements within that GO category or just because that’s where these elements are typically found. You will have to do your own enrichment analysis by accounting for this distribution of elements. If you take 1000 random subsample (same number as you have of accelerated elements) of all conserved elements and then take your accelerated elements and compare this to the distribution, a GO category should be marked as enriched if it is significantly different from the background distribution.
Enhancer activity
> I found this section to be quite interesting and important to the overall strength of the paper. As this is not my area of specialty, I digress from making comments about the methodology.

Other comments:
> At the beginning of your results section are details that should go into the methods. Like it is important to know that, “in our search for accelerated regions in mammals, our approach required that the most basal member of the mammals, the platypus (Ornithorhynchus anatinus), must be present in all alignments and must share a common nucleotide change with all other mammals” in the methods and not in the results section.
155,630 conserved sequences vs 93,881 conserved
> The 100 way and 77 way alignments with only a few changes and many shared taxa revealed a large discrepancy of conserved elements. Since both of these are tetrapod conserved elements, what is the overlap between these two numbers?

Reviewer #2

(Remarks to the Author)

The authors have performed a comparative genomic analysis of mammals and birds to find regions of accelerated evolution. They found both coding and non-coding accelerated elements on both the mammalian and avian lineages. They further perform GO category enrichment to see which kind of genes are likely to be affected by the accelerated elements. They find that some elements overlap between different accelerated data sets (i.e. between species clades)

P6: The authors write: “In mammals, accelerated sequences were found to be almost twice as likely to be coding (20,531 out of 24,007; 85.6%) as they were in birds (2,771 out of 5,659; 49%). In contrast, in birds, the proportion of conserved noncoding sequences over the total number of conserved elements is 51%, whereas in mammals, the same proportion is 14.4%.” Could this be an artifact because a: gene annotation is not so good in birds as in humans? b: the evolutionary distance to platypus is so long that it primarily captures coding elements?

In conjunction with this: What is the evolutionary time spanned in the mammalian and avian analysis?

How do the results change if platypus is not included in the analysis?

P 10: “The three regions showing the highest frequency of ncAvAR were: chr2:90246100-91807271 (a gene desert according to G. gallus gene annotations” What are the nearest gene on each side?

P13: Do you lose information/elements when transferring accelerated regions to the human genome for the pathway analysis?

P17 “To analyze this, we cross-referenced mammalian and avian noncoding accelerated regions, which led to the identification of 285 common accelerated noncoding regions (CARs) (Supp. Table 14).” Is this a statistically significant enriched overlap?

P34 “When necessary for some tools, we converted the accelerated chicken (Gallus gallus) galGal6 elements to hg38 genomic coordinates using liftOver tools (https://master.bioconductor.org/packages/release/workflows/html/liftOver.html).” What fraction of elements lifted over?

Figure 1: mark up scale for the two trees either as time or as substitutions per site

Figure 2 c,d add unit on x-axis

Reviewer #3

(Remarks to the Author)
In this manuscript, the authors perform a comparative analysis of different animal genomes looking for genomic regions that accumulate more changes than expected, according to the normal mutation rate. Using this methodology, they identify sets of highly accelerated regions specific for mammals and birds. They argue that the group of accelerated regions presented in this study for mammals is more accurate than the one previously published, thanks to the introduction of platypus in the genomic comparison in this work, while only therian species were used previously.
Some of the accelerated regions identified are tested for regulatory activity using transgenic assays in zebrafish. The authors claim that both, human versions of these putative regulatory regions and their orthologous sequence in alligator (taken as an outgroup of mammals) drive the reporter expression to similar domains, despite the good number of specific changes accumulated in mammals in these regions. They attribute this to compensatory changes in these sequences that maintain a functional conservation, although the sequence conservation is partially lost. This interpretation sounds too complicated to me, when the similarity among different founders’ expression in some of the lines is controversial. The selected regions drive expression to very broad domains, including neural tissue, and this makes very difficult to evaluate whether the expression driven by two homologous regions (or even different insertions of the same region) are equivalent. Furthermore, I find that this statement is not essential for the whole story, just probing that these regions act as enhancers is the important message, since it demonstrates their capacity to contribute to the emergence of specific innovations. In addition, I think that these two messages are somehow contradictory: 1) accelerated regions are “critical for the evolution of specific traits”; and 2) mammalian and reptile orthologous regions drive similar expression patterns, which suggests that “some substitutions were probably fixed to compensate for other changes maintaining the expression pattern unaltered”.
In my opinion, one of the most interesting conclusions of the manuscript is that accelerated regions have contributed to the emergence of specific novelties of each animal group. However, this conclusion is not supported by appropriate statistical analysis and figures, it is just based on the direct observation of GO terms enriched in genes associated to these genomic regions (for example: “This analysis showed us that those genes with clustering of ncMARs are more frequently related to typical mammalian traits such as lactation, diaphragm changes and dentition”). In order to support this and other conclusions, I would suggest thinking of a statistical analysis where the enrichment in GO terms specific for evolutionary novelties were compared to the enrichment in other terms related to ancestral traits, if such analysis is possible.
On the other hand, some unexpected GO terms are ignored, like ‘lactation behaviour’ in the shared accelerated regions between mammals and birds. Besides, results from GO term enrichment analyses could be presented in some kind of friendly plots, like dotplots, in which some important values can also be shown (number of genes selected vs. genes in the category, pvalue, etc.).

Strengths of the work:
- The general approach and the methodologies used are correct and the conclusions of the study are interesting.
- Introducing a monotreme in the mammalian comparative is definitely useful in order to identify genomic characteristics present in the mammal last common ancestor.
- Performing functional assays help to validate the results, probing that accelerated regions have indeed regulatory activity.

Weaknesses of the work:
- The approach is not especially innovative.
- Language use is not good enough, it requires a strong review.
- Figures are repetitive and there are important conclusions that are not supported by them.
- Lack of statistical analyses that support some conclusions.

Taking all this into account, the manuscript should be strongly revised to deserve its publication in Nature Communication. As a summary, I will give a general outline of the main changes that should be made, in my opinion:
- Revision of language in general.
- Move some of the main figures regarding to transgenesis assays to supplementary info.
- Generate some new figure(s) with GO term enrichments as dotplots (or similar) and statistics.
- Perform a statistical comparative analysis to support GO-based conclusions, like the involvement of accelerated regions in the emergence of evolutionary novelties.
- Explain the contradiction of accelerated regions maintaining their function but contributing to novelties, or revisit some of these interpretations.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Review of Comparative genomics sheds light on mammalian and avian gene regulation and phenotypic evolution
First of all I would like to acknowledge the extensive work done by the authors to revise the manuscript based on the three reviewer comments with detailed response to each point raised. This has definitely improved the manuscript a lot.
Fig1:
Upon going through the manuscript again, one thing that caught my eye which I did not see on the first time around is the bird tree in Fig 1. That tree is not the correct phylogenetic relationships of birds. The first branch is correct with Struthio and Tinamus. But then according to this tree, the next branch is Picoides, a woodpecker. Then there are two clades, one containing a hodgepodge of random birds and a second that in some respects looks like the correct coarse relationships among birds but with some glaring issues as well. The woodpecker is not the earliest branching Neoaves. There is no support for this two group branching in literature. The galloanserae (chicken and geese) are embedded somewhere within group 1 and broken up and the galliformes are sister to an owl (which itself is separated from other birds of prey in clade 2). Also, in clade 2, the falcons are sister to birds of prey (not to parrots). The rifleman is not the earliest branch of the passerine birds (instead it is the ground jay). Where did you get this tree? How did this tree happen?
Second, minor point, the branch labelled mammals on the figure is in the wrong spot (should be one more node down, currently the highlighted branch is the reptile branch).

Methods:
With regards to the methodology, thank you for clarifying your methods more and providing a more detailed description of the code and methods. It makes more sense now as to how you are defining acceleration. Also, yes the paucity of alignments containing all amniotes warrants the use of the 100 way and 77 way alignments and even though these alignments may have issues due to the quality of samples/genomes, it hopefully does not affect conserved regions as much.
Minor point: in the methods the common names of species are inconsistently written down with uppercase and lowercase names (e.g. Hooded Crow vs american eagle). I would advise just standardizing the names based on the journal recommendations.
Also, the gBGC analysis helps clarify some of these accelerated elements. Thank you for taking the time to address that.

Results:
Line 127: I would change the word basal with earliest diverging. The word ‘basal’ has been controversial since these taxa are still currently living and not ‘basal’ in the traditional sense of the word. You can have a basal node (the node where the paleognaths and neognaths split) but not a basal bird (unless you literally mean the bird that lived 170 million years ago).
On a related note. I suspect you dropped from 155,630 elements to 5,659 elements in the bird dataset because the tinamous seems to have a highly accelerated rate of evolution, the often appear in this long branch on the avian tree and so many more fixed differences are probably in the tinamous lineage. My guess is that if you made your criteria as only applying to ostrich, you’d find more shared conserved elements that meet your criteria. This also explain the discrepancy you address for reviewer 2. Just worth a check.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have properly addressed the reviewers comments.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The manuscript has been strongly revised following the reviewers suggestions, and highly improved. My concerns have been addressed.
Therefore, I endorse it for publication in Nature Communications.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Thank you for making the changes.

While that tree suggests there is something off with at least some of the sites in the alignment, I suspect that is due to paralogy and other issues. I agree with the authors that since they are testing the earliest bird branch that the order of nodes at the other parts of the tree should not affect the results as much since they are looking for sites that are conserved in the recent branches and accelerated at the base node (which is correctly placed).

I am satisfied with the current draft of the manuscript.

(Remarks on code availability)

RESPONSE TO REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author)

Corrections suggested by this reviewer are highlighted in yellow in the revised manuscript.

While I found the premise and the general scope of the results pretty interesting, I found it hard to follow the methods used to generate the datasets and so have several questions as to what exactly was done. While there were code associated with the methods, they were not easily interpretable, particularly the early element detection and filtering steps (steps on Github that is not in the Rmarkdown file). I shall reserve final judgements on some of the results and discussion pending more clarification of the methodology.

Comment (C)1. My first major question is this:
Why do you expect accelerated elements to be functional or functionally relevant? My interpretation of the logic is that conserved elements are what are important since conservation pertains to potential functional relevance (as transcriptional factors/binding sites). Acceleration is just the lack of conservation, and it may or may not still have any functional relevance. What you can infer from acceleration is that something functional that was conserved across time is not conserved anymore (hence probably nonfunctional), it does not necessarily mean that it has adapted to a new function. If an element is really functionally relevant then you will may see rapid acceleration at the base of a lineage (say all birds) but then subsequent conservation (since they are functional). I think you are looking for things that are accelerated and have stayed that way across the family (unless I am mistaken). Also, conserved regions tend to move around as motifs for binding change, new regions evolve. So, loss of conservation does not necessarily mean that there has been functional change. It could easily mean that the constrained region has moved somewhere else.

Response (R) 1. Thank you for your insightful comments. We understand your point regarding the functional relevance of accelerated elements. As you pointed out, if an element is functionally relevant within a specific lineage (e.g., all birds), we would expect to see rapid acceleration at the base of the lineage followed by conservation and this is exactly what we are seeing. Our approach is designed specifically to find this kind of element that underwent acceleration in the base of a lineage (either mammals or birds) and then subsequent conservation along this lineage. More precisely, our candidates accelerated regions were identified by intersecting conserved elements in the outgroup (amniotes alignment) with conserved elements in the ingroup (mammals or birds). This approach allows us to infer functional

relevance, as these regions exhibit conservation across both groups, albeit an elevated rate of nucleotide change compared to neutral regions between groups. This suggests that while traditional conservation may have been lost or modified, the regions retain a form of conservation within the lineages analyzed. In other words, acceleration is observed within one lineage, but differential conservation is maintained between groups, which is consistent with functional relevance despite evolutionary change.

To improve the clarity and reusability of our code, we rewrote the accelerated elements computation as a standalone package (https://github.com/paulati/cladeAcc). The same code, with different parameters, can now be used to compute accelerated elements for both mammals and aves. We are addressing code-related questions by referring to both the original implementation and the updated version.

To increase methodological clarity, we share here links to the GitHub page of the specific scripts we used to generate the intersections of conserved regions in mammals and birds:

In the original version:

- **Mammalian conservation intersection: mammals/conservation_intersection_mammals_sarcopterygii.R**
- **Avian conservation intersection: aves/conservation_intersection_aves_sarcopterygii.R**

In the new version:

function conserved_elements_in_common at https://github.com/paulati/cladeAcc/blob/master/R/conservation.R

These conservation intersections define the regions where we analyzed acceleration, capturing those elements that, although may have altered their evolutionary rate, still show conservation in a way that suggests potential functional relevance in the vertebrate groups analyzed.

In addition, regarding functional relevance, previous studies, including work from our own lab and others in the field, have indeed shown functional changes associated with accelerated regions (Kamm et al. 2013; Berasain et al. 2024; Caporale, Gonda, and Franchini 2019; Caporale et al. 2024; Boyd et al. 2015; Whalen and Pollard 2022; Whalen et al. 2023)**. In the present work we analyzed selected elements´s function as transcriptional enhancers. We assayed the top five ncMAR sequences (ncMAR1 to 7; ncMAR2 was not included because it shows epigenetic signatures of promoter function and ncMAR6 intersects a pseudogene) and found that all of them behave as strong transcriptional enhancers during development in transgenic zebrafish assays (Fig. 4 and Supp. Figs. 3 to 7). We also analyzed four of them comparatively testing human (mammal) and alligator (non mammalian vertebrate) sequences. None of them showed functional differences in temporal or spatial expression pattern at the stages analyzed, even though both sequences tested display numerous nucleotide differences. In addition, we also tested four sequences that are at the same time**

accelerated in the lineage leading to mammals (ncMARs) and in the lineage leading to humans (HARs) located in the gene *NPAS3* (Fig. 5 and Supp. Figs. 8 to 13) and one sequence that shows signatures of acceleration in the basal Aves lineage and in the basal Mammals lineage. In these cases, we found differences in expression in two of them, displaying either gain of function (HAR173/ncMAR506) or loss of function (ncMAR3220/ncAvAR1232) in the mammalian lineage. Our results suggest that some substitutions were probably fixed to compensate for other changes maintaining the expression pattern unaltered whereas others were probably selected because they induce a change in regulatory function. Our data coincides with previous results showing that acceleration can induce expression changes in around 30 to 50 % of the accelerated sequences ((Berasain et al. 2024; Whalen et al. 2023; Capra, Erwin, et al. 2013; Kamm et al. 2013; Uebbing et al. 2021; Caporale, Gonda, and Franchini 2019)**.**

C2. With that, I will present some of the methodological questions that I had based on what you have in your methods section:
“Because there were no multiple alignments that included a large number of mammals, birds, and other amniotes at the time of this study”
> While I can see that based on the dates on the Rmarkdown file much of the results were generated and analyzed in 2020, this statement then was probably valid. However, in the subsequent four years many new alignments have been produced. While I don’t advocate for the authors to use these new datasets, I still think some of the results from these studies must be acknowledged. In particular, the introduction should be updated with some of the studies that were derived from the Zoonomia or the B10K alignments to address convergence. I would probably include a year in that sentence above to allow readers to interpret time of study as 2020 and not 2024/5 when this paper would be published.

R2. We thank the reviewer for this comment because it allows us to clarify this important point. The Zoonomia project recently published (Christmas et al. 2023) **includes in its alignment only placental mammalian species, and it lacks non-mammalian vertebrates like birds, reptiles, and even marsupials and monotremes such as the platypus, which are essential for our study's comparative approach. On the other hand, the B10K project is an initiative to generate representative draft genome sequences from all extant bird species and includes to date 363 bird species** (Stiller et al. 2024; Zhang et al. 2015)**, however, the available alignments only include avian species.**
Our study aims to identify vertebrate conserved regions that underwent acceleration specifically in either mammals or birds basal lineages by comparing these groups with a representative and brader set of amniotes. Thus, the statement included in our previous version, "there were no multiple alignments that included a large number of mammals, birds, and other amniotes at the time of this study," remains accurate, as no existing alignment dataset, including Zoonomia, meets the specific requirements of our analysis. Nevertheless, to acknowledge the reviewer's suggestion, we mentioned Zoonomia and B10K projects in the Introduction and explained their contributions to mammalian and birds comparative genomics (pages 3 and 4). We also mentioned that, due to the absence of non-mammalian species in Zoonomia and the absence of other vertebrates in the B10K alignment, they cannot be directly used in our study, which requires alignments across both mammalian and avian lineages

alongside other amniotes. There is an available 605 vertebrate species (https://cgl.gi.ucsc.edu/data/cactus/605-vertebrate-2020-hub/hub.txt), however, this alignment does not include non-placental mammals such as monotremes and marsupials, that are key to our analyses. Thus, we also rephrased the text to clarify even further why we chose to use the 100 pathways, https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/) and 77 pathways, http://hgdownload.cse.ucsc.edu/goldenPath/galGal6/multiz77way/) alignments.
The sentence include in page 29 now reads:
"At the time we started this study (2020) there were premade multiple alignments that included a large number of mammals, birds and other amniotes, so we decided to conduct our search using the intersection of two large independent multiple alignments: one containing primarily mammals along with other amniotes ("100 pathways", https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/) and another one containing primarily birds and other amniotes ("77 pathways", http://hgdownload.cse.ucsc.edu/goldenPath/galGal6/multiz77way/)".

C3. "Then, first, we identified conserved regions in amniotes of at least 50 bp using phastCons software"
C3.a I can see from the code that you have generated and used a neutral rate file to use for estimation of conserved elements. However, the information of what you actually used to generate this neutral rate file is absent. What kind of marker was used for this? Or maybe you got this file from somewhere else? Please include details in the methods.

R3. Thanks for this comment that allows us to clarify this important matter. For both, the calculation of conserved elements and accelerated regions we used the same neutral model, a continuous Markov model of nucleotide substitutions derived from neutral sequences. Here we use the REV model fitted to 4-fold degenerate sites. In mammals, we used the neutral model file `hg38.phastCons100way.mod`. This file was downloaded from the UCSC Genome Browser repository (https://hgdownload.cse.ucsc.edu/goldenpath/hg38/phastCons100way/hg38.phastCons100way.mod) and used to estimate conservation in mammalian regions.
In the original code, it is referenced on line 35 of the configuration file https://github.com/paulati/acc_regions_mammals_aves/blob/main/mammals/common/config/paths_config_modneuPhastCons100way.R
In the new version, it is referenced on line 9 of the configuration file https://github.com/paulati/cladeAcc/blob/master/inst/config.yml.
For birds, we used the neutral model `galGal6.phastCons77way.mod`. This file was also obtained from the UCSC Genome Browser (https://hgdownload.soe.ucsc.edu/goldenPath/galGal6/phastCons77way/galGal6.phastCons77way.mod) and was specifically used for avian regions. In the original code, it is referenced on line 14 of the configuration file (https://github.com/paulati/acc_regions_mammals_aves/blob/main/aves/r_source/config/config_common.yaml).
In the new version, it is referenced on line 13 of the configuration file https://github.com/paulati/cladeAcc/blob/master/inst/config.yml

In this revised version we include these details in the Methods section under a new subsection (Neutral Model, page 28) clarifying the sources and configurations we used to generate the neutral rate files of our conservation analyses.

C3.b. How did you generate the conserved rates?

We calculated conservation scores using the rphast library (http://compgen.cshl.edu/rphast/rphast-manual.pdf). Instead of analyzing the conservation score directly, we used the most.conserved attribute from the output of the phastCons function. According to the rphast documentation, most.conserved is "an object of type feat which describes conserved elements detected by the Viterbi algorithm." This output provided the conserved elements, which we then intersected with regions of the alignment that met our criteria for informative regions (i.e., regions with the required species in each group).

For the avian analysis, this process is implemented in

* **Original code: lines 151 to 157 of conservation_aves.R and in lines 134 to 140 of conservation_sarcopterygii.R. For the mammalian analysis, this process is carried out in lines 63 to 69 of conservation_base.R.**

* **New version: lines 10 to 18 of conservation.R**

We have incorporated these details into the Methods section (pages 29 to 32) to clarify our approach for generating conserved regions.

C3.c. Scientific names are not italicized

Thank you for pointing this out. We have reviewed the manuscript and ensured that all scientific names are now italicized.

C4. "A region was considered informative when at least three outgroup species, i.e., those that are not part of the set where acceleration is sought (in this case non-mammalian amniotes) were present"
> Only by looking at the Github page and looking at parts of the code was I able to figure out that you may have grouped the turtles as one (since you sampled four turtles); or maybe I am wrong? Otherwise, if you choose three outgroups how did you make sure all three were not turtles? That would definitely bias your results to turtle conserved elements. Again, I am assuming you did account for this but what you did is not mentioned in the methods.
R4. Thanks for this comment that allows us to clarify this point. The reviewer´s interpretation is correct. There are four turtle species in the alignment, but we require only one to be present for a region to be included.
Specifically, in the original code, line 54 in the script (https://github.com/paulati/acc_regions_mammals_aves/blob/main/mammals/conservation/scripts/conservation/conservation_sarcopterygii.R) sets `min.numspec=1` for turtles, meaning that only one turtle species needs to be present.
Furthermore, to prevent the results from being biased towards turtle-conserved elements, line 50 in the same file specifies that two particular outgroup species, *Alligator mississippiensis* (allMis1) and *Anolis carolinensis* (anoCar2), must also be

present (`min.numspec=2`). Line 60 intersects these criteria, ensuring that the selected regions are considered informative only if they include at least one turtle along with both *Alligator mississippiensis and Anolis carolinensis*.
In the new version of the code, this same approach is implemented in the function required_species_features_sarcopterygii_100way at https://github.com/paulati/cladeAcc/blob/master/R/conservation_delegates.R
This approach minimizes any potential bias from over-representation of turtles and ensures that the regions reflect broader conservation across amniotes.

In this revised version we have updated the methods section (Conserved elements computation; page 31) to clarify these criteria for selecting informative regions.

C5. “we defined accelerated regions in a particular lineage as those regions that exhibit acceleration signals within a functional element”
> I am confused by this, by this do you mean you are choosing accelerated elements as any element that is accelerated in any or combination of mammalian species you selected? Does it mean an element that is conserved in all species and say accelerated in primates would be classified as accelerated?

R5. Thanks for pointing out this phrase that is not correctly elaborated. We understand that the confusion may stem from our use of the term "functional element," which could imply known functionality. In our case, we refer to these as "functional candidate elements" due to their conservation across both groups, though we do not have evidence of specific functionality. We have corrected the use of this term across the manuscript.
We do not mean that we are choosing accelerated elements as any element that is accelerated in any or combination of mammalian species selected. In the case of mammalian accelerated regions or MARs the signature of acceleration is always in the basal branch leading to mammals (Fig. 1, d).
Since we understand that the description of the methods to identify accelerated sequences raise some confusion, we have edited this description in pages 32 to 34 (Accelerated Regions Computation) in the Methods section to clarify it.

C6. Bird acceleration
> With regards to birds, I think it is important to account for gc-biased gene conversion as well. Since, an element with high GC content can be flagged as accelerated due to gc-biased gene conversion and not due to any selective pressures.
o Liu, Anguo, et al. "GC-biased gene conversion drives accelerated evolution of ultraconserved elements in mammalian and avian genomes." Genome Research 33.10 (2023): 1673-1689.
R6. Thanks for your suggestion. We followed your recommendation and to account for gc-biased gene conversion (gBGC), we used the same methodology of (Liu et al. 2023) **and ran the software phastBias** (Capra, Hubisz, et al. 2013) **to detect gBGC tracts in the accelerated elements. We were able to detect 276 phastBias tracts in 182 ncMARs accelerated elements. We did the same for ncAvARs and found that 37 phastBias tracts intersect 31 avian accelerated elements. These results included in**

Supp. Tables 6 and 7 show that gBGC affected a minimum proportion of the identified accelerated elements. We incorporated these new results into the manuscript in pages 6 and 7. We also described in detail the method used in the Methods section under the subtitle "GC-biased gene conversion detection" (pages 34 and 35).

C7. GO analysis

> Using GO analysis without taking into account the distribution of elements will give you a skewed result. The base assumption of using the GO analysis is that conserved elements are randomly distributed throughout the genome and hence you expect a random set of background genes. However, this is not true. We know conserved elements are more common near specific genes and gene regions. And a few genes have large number of conserved elements nearby. Here is an example, but I think there is another paper that shows this to be true in birds and mammals as well.

o Gonzalez, Paul, Quinn C. Hauck, and Andreas D. Baxevanis. "Conserved Noncoding Elements Evolve Around the Same Genes Throughout Metazoan Evolution." Genome Biology and Evolution 16.4 (2024): evae052.

So, if you take a biased set of elements, you are more likely to encounter elements that are accelerated in GO categories that are already overrepresented by conserved elements. Hence, you cannot make any inference as to whether you see a certain GO category as being enriched is due to the prevalence of large number of accelerated elements within that GO category or just because that's where these elements are typically found. You will have to do your own enrichment analysis by accounting for this distribution of elements. If you take 1000 random subsample (same number as you have of accelerated elements) of all conserved elements and then take your accelerated elements and compare this to the distribution, a GO category should be marked as enriched if it is significantly different from the background distribution.

R7. Thank you for this insightful comment. We understand the concern regarding the potential bias in GO analysis due to the non-random distribution of conserved elements around specific genes, as well as the risk of overrepresentation in certain GO categories. Following your suggestion in the new version we implemented a new strategy to analyze GO terms using rGREAT (Gu and Hübschmann 2023)**. Using this approach we compared the results of rGREAT obtained from the set of accelerated regions with those obtained from equivalent sets of conserved regions. To do that we sampled conserved regions, building 5,000 sets with the same size and chromosome distribution as the accelerated regions set, and evaluated rGREAT on each of these sets. First, we built an empirical distribution of 'n_hits' values for each specific go term from rGREAT results of each of the 5,000 conserved regions subsets. Then we calculated the 'n_hits' value from the rGREAT result of the accelerated regions set (acc_observed_region_hits). Finally, we calculated the proportion of random n_hits observed that are greater than acc_observed_region_hits. If this proportion was less than 0.05, then this set of accelerated regions differs from the random sets of conserved regions evaluating this specific GO term, and we can assume that the behavior of the accelerated regions set is independent of its conserved nature for this term. We considered this proportion as an empirical p-value, and this p-value for each GO term was corrected for multiple comparisons. The Benjamini-Hochberg method was applied to compute approximate false discovery rates (FDRs) for each term.**
We report a GO term as associated with the accelerated regions set if:

1. The behavior of the accelerated regions set is independent of its conserved nature and
2. The adjusted p-value returned by the binomial test performed by rGREAT on the accelerated regions set is less than 0.05.

We have updated the methods section **Gene ontology analysis** (page 35 to 37) and we provide the code and more details of the implementation at (https://github.com/paulati/ncomms-24-47757/tree/main/go_analysis). The results are described in pages 11 to 13 and included in Supp. Tables 14 to 19, Fig. 2 and Supp. Fig. 3)

C8. Enhancer activity
> I found this section to be quite interesting and important to the overall strength of the paper. As this is not my area of specialty, I digress from making comments about the methodology.
R8. Thanks for your comment, we appreciate your consideration of our functional experiments.
C9. Other comments:
> At the beginning of your results section are details that should go into the methods. Like it is important to know that, “in our search for accelerated regions in mammals, our approach required that the most basal member of the mammals, the platypus (Ornithorhynchus anatinus), must be present in all alignments and must share a common nucleotide change with all other mammals” in the methods and not in the results section.
R9. Thanks for pointing this out, however, we consider it important to make this statement at this point since the presence of Platypus in our analysis makes an important difference compared to approaches followed in the past by other researchers.

C10. 155,630 conserved sequences vs 93,881 conserved
> The 100 way and 77 way alignments with only a few changes and many shared taxa revealed a large discrepancy of conserved elements. Since both of these are tetrapod conserved elements, what is the overlap between these two numbers?

R10. Thank you for your comment. We used these two alignments because they do not share many species since the 100 way is mainly composed of mammals (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/multiz100way/) and the 77 way is mainly composed of birds (http://hgdownload.cse.ucsc.edu/goldenPath/galGal6/multiz77way/)
We found that 31 species overlap between the two alignments (25 when considering exact genome versions). Since we used a subset of species from the 100-way and the 77-way alignments, in the final alignments used, only seven species overlapped. That explains the discrepancy between the number of conserved elements.
Additionally and following your suggestion, we compared the overlap between the two sets of conserved elements (155,630 conserved sequences in birds and 93,881 conserved sequences in mammals). We did a reciprocal liftOver and compared the overlap of conserved regions. When the 155,630 bird conserved sequences were lifted over from the chicken genome (galGal6) to the human genome (hg38), we could map 108,968 conserved regions in birds that overlap with 68,703 conserved regions in

mammals (73% overlap with mammalian conserved elements). These overlap percentages remain consistent when performing the reciprocal liftover (hg38 to galGal6).

Reviewer #2 (Remarks to the Author)

Corrections suggested by this reviewer are highlighted in green in the revised manuscript.

The authors have performed a comparative genomic analysis of mammals and birds to find regions of accelerated evolution. They found both coding and non-coding accelerated elements on both the mammalian and avian lineages. They further perform GO category enrichment to see which kind of genes are likely to be affected by the accelerated elements. They find that some elements overlap between different accelerated data sets (i.e. between species clades)

C1 (P6) The authors write: “In mammals, accelerated sequences were found to be almost twice as likely to be coding (20,531 out of 24,007; 85.6%) as they were in birds (2,771 out of 5,659; 49%). In contrast, in birds, the proportion of conserved noncoding sequences over the total number of conserved elements is 51%, whereas in mammals, the same proportion is 14.4%.”
Could this be an artifact because

a: gene annotation is not so good in birds as in humans?

b: the evolutionary distance to platypus is so long that it primarily captures coding elements?

How do the results change if platypus is not included in the analysis?

R1. Thank you for your questions and comments that allow us to clarify this point. Regarding the differences in the proportions of coding versus non-coding accelerated elements between birds and mammals, these findings are consistent with previous studies, such as (Seki et al. 2017)**. This study identified avian-specific highly conserved elements (ASHCEs), which predominantly (over 99%) reside in non-coding regions and are suggested to play significant regulatory roles, especially in the evolution of bird-specific traits. It further supports the hypothesis that non-coding regulatory elements are a critical driver of avian evolution, contrasting with mammals, where lineage-specific coding innovations have played a larger role** (Kaessmann 2010)**.**

a: gene annotation is not so good in birds as in humans?

We have considered the potential impact of gene annotation quality on the observed proportions of coding versus non-coding accelerated elements in birds and mammals. Park et al. (2023) evaluated gene annotation quality across 114 species, including *Gallus gallus* and *Homo sapiens*. This study highlighted that there are different annotation strategies of protein-coding genes, where they are more frequently annotated in birds (*Gallus gallus:* 56.487% of annotated genes), and that this proportion is lower in the human genome (33.041%). Interestingly, this implies that bird annotations are biased towards conservative protein-coding regions. If annotation quality were to cause artifacts, the opposite pattern to our observation (a higher coding fraction in mammals) would likely emerge.

b: the evolutionary distance to platypus is so long that it primarily captures coding elements?

To address this question and the question about how the proportions of coding versus non-coding accelerated elements change when platypus is excluded from the analysis we would need to analyze datasets focused on eutherians that exclude monotremes, such as those provided by Holloway et al. 2016 or Zoonomia (Holloway et al. 2016; Christmas et al. 2023)**. If the evolutionary distance to platypus significantly impacts our non-coding detections, we would expect that excluding platypus would influence the proportions of coding and non-coding elements. It was not within the scope of this study to calculate accelerated elements specifically for eutherian mammals, as these have already been extensively reported in the previously mentioned studies. Therefore we analyzed the proportion of coding and noncoding elements in conserved and accelerated elements in mammals in other studies. In particular, Holloway et al. 2016 reports:**

"*177,346 vertebrate genomic regions that are conserved among therians (therian conserved regions; TCRs), of which 4,797 have a strong signature for accelerated evolution in the therian ancestor (TSARs). [...] The distribution of TSARs with respect to human gene features is similar to that of TCRs" [...]. More than half of the TSAR sequences overlap exonic regions of the human genome.*."

3,175 out of *4,797 (66%)* TSARs overlap exonic regions "Table S1. TSAR annotation" (Holloway et al. 2016)**.**

These previous results align with our observation that a larger proportion of accelerated elements in mammals are coding compared to noncoding indicating that excluding the platypus does not influence the trend.

In conjunction with this: What is the evolutionary time spanned in the mammalian and avian analysis?

Considering previous studies the mammalian lineage has been calculated to emerge around 170 mya (Bininda-Emonds et al. 2007) **and the bird lineage has been estimated to appear around 165–150 million years ago from theropod dinosaurs during the Jurassic period** (Zhang et al. 2014)**.**

Taking into consideration your suggestion, we have added a Supp. Table with these numbers and comparisons (Supp. Table 5) and reworded the paragraph in the manuscript in the Results section to add more clarity, that nows reads as follows:
"We found striking differences between the proportions of noncoding and coding accelerated sequences in mammals and birds (Fig. 1, C, Supp. Table 5). In mammals, 85.6% of accelerated elements (20,531 out of 24,007) and 78 % of base pairs (4,261,915 of 5,449,351 bp) were coding, while only 14.4% (3,476 out of 24,007) covering 1,187,436 bp (22 % of total) were noncoding. In contrast, birds showed nearly equal proportions of coding and noncoding accelerated elements, with 49% of elements (2,771 out of 5,659) and 900,855 bp (of 1,981,612 bp) being coding and 51% (2,888 out of 5,659) including 1,080,757 bp being noncoding (Supp. Table 5). However, it is important to note that these proportions follow a trend already observed in the proportions of conserved coding and noncoding regions in our mammalian and bird alignments (Supp. Table 5). These data suggest that accelerated evolution is shaping

these two functional components of the genome, coding and noncoding, with different intensities in mammals and birds."
In addition, taking into consideration your comments we have further discussed the matter in the Discussion section in pages 21 and 22.
"Our findings reveal a distinctive difference in the proportion of noncoding accelerated elements between mammals (3,476/24,007; 14.4%) and birds (2,888/5,659; 51%), suggesting that the role of accelerated evolution in shaping genome function differs between these two vertebrate groups. It was previously found that most TSARs were located in coding regions and these coding accelerated regions show high rates of nonsynonymous (dN) substitution compared to synonymous (dS) substitution in the branch of the therian ancestor, which may indicate adaptive evolution of the proteins involved (Holloway et al. 2016)**. These and our results suggest that in mammals, coding evolution appears to play a key role and is readily detectable with accelerated evolution studies. In contrast, studies of avian genome evolution have found that functional noncoding regions are more plastic and less strongly conserved than coding regions are more frequently subject to acceleration** (Feng et al. 2020)**."**

C2 (P10): "The three regions showing the highest frequency of ncAvAR were: chr2:90246100-91807271 (a gene desert according to G. gallus gene annotations" What are the nearest gene on each side?

R2. We intersected the mentioned TAD region with the *G. gallus* ncbiRefSeq All gene annotations, including both curated genes and computationally predicted transcripts. The results are:

TAD coordinates	Inside Genes	Closest Genes
chr2:90246100-91807271	*LOC112531824, ZNF516, TSHZ1, ZNF407, TDP2, GALR1, MBP, ZNF236, LOC107052677, ZADH2, LOC112531991*	*CNDP1, ACOT13*

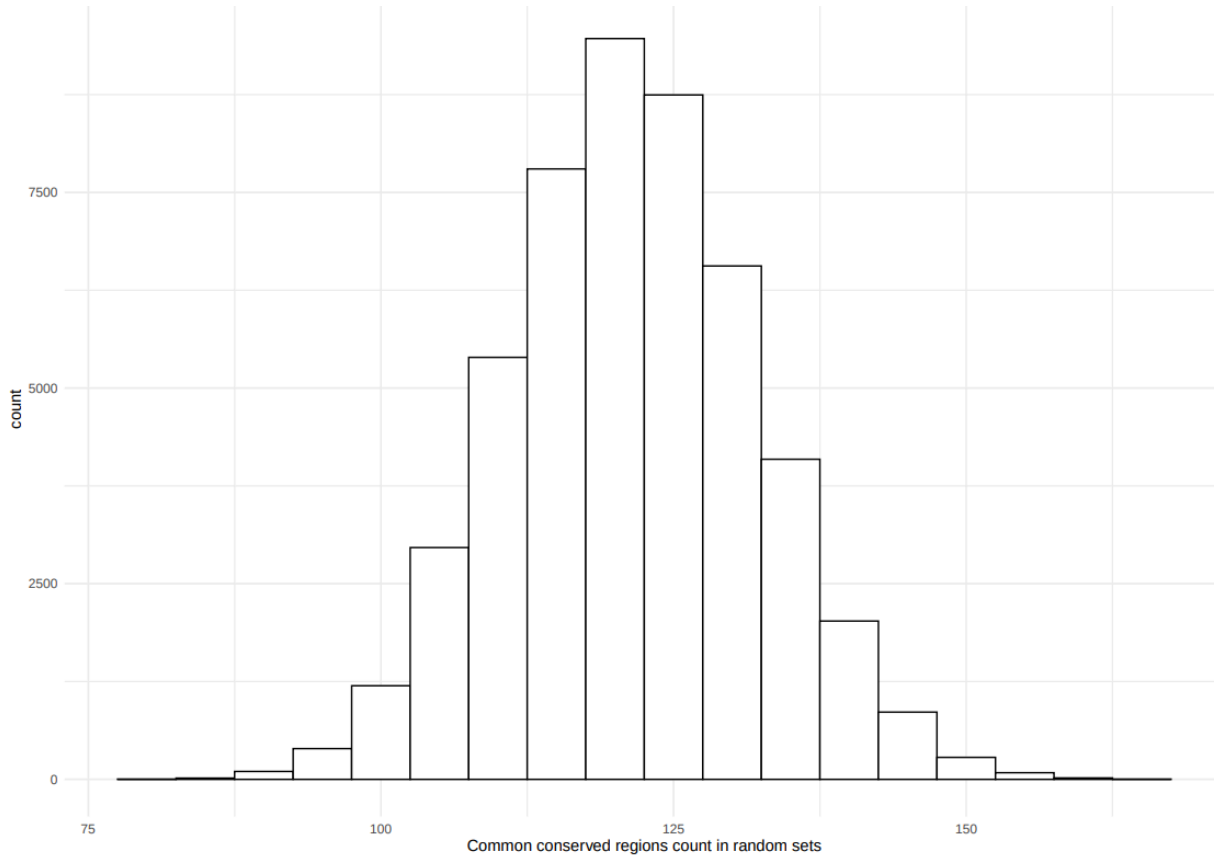
C3 (P13): Do you lose information/elements when transferring accelerated regions to the human genome for the pathway analysis?

R3. Some information is lost when transferring avian accelerated regions to the human genome: 584 out of 2888 ncAvAR regions (about 20.2%) could not be mapped. However, taking into account this Reviewer and Reviewer1´s comments, we modified some of the analysis so that they are now performed directly in the *Gallus gallus* genome in order not to lose elements. For instance, in this new version part of the analysis of Signatures of regulatory activity in noncoding accelerated elements (page 13 in the Results section) was performed using the chicken-specific regulatory data from the FAANG project (Pan et al. 2023) **available at**

(https://farm.cse.ucdavis.edu/~zhypan/Chicken_FAANG/Tissue_specific_regulator/) (see Results in pages 13 and 14).

C4 (P17): “To analyze this, we cross-referenced mammalian and avian noncoding accelerated regions, which led to the identification of 285 common accelerated noncoding regions (CARs) (Supp. Table 14).” Is this a statistically significant enriched overlap?

R4. We apologize but we don't exactly understand this question. This approach was simply aiming to identify MARs and AvARs regions that overlap in the genome. Thus, these regions that we named Common Accelerated Regions (CARs) underwent acceleration in the lineage leading to birds and in the lineage leading to mammals. To answer the question and to assess whether the overlap of noncoding accelerated regions between mammals and birds represents a statistically significant enrichment, we compared the number of overlapping regions between randomly generated sets of conserved regions in these lineages. We built 50,000 random samples of conserved regions, matching the number and chromosomal distribution of MARs and AvARs. For each pair of random sets, we calculated the number of overlapping regions. The distribution of intersection counts is shown in the plot below. We then calculated the empirical p-value for the observed overlap of 285 (the number of CARs) as (number of intersections ≥ 285 + 1) / (number of samples + 1). The resulting empirical p-value was less than 2e-5. Thus, we can conclude that the number of MARs and AvARs common regions is significantly greater than expected from intersections of random conserved sets in mammals and aves. The implementation of this strategy is available at https://github.com/paulati/ncomms-24-47757/blob/main/car/source/car_significance.R



C5. P34 “When necessary for some tools, we converted the accelerated chicken (Gallus gallus) galGal6 elements to hg38 genomic coordinates using liftOver tools (https://master.bioconductor.org/packages/release/workflows/html/liftOver.html).” What fraction of elements lifted over?

R5. This was replied to in C3 (P13). As mentioned, some information is lost when transferring avian accelerated regions to the human genome: 584 out of 2888 ncAvAr regions (about 20.2%) could not be mapped. However, taking into account this Reviewer and Reviewer1´s comments, we modified some of the analysis so that they are now performed directly in the *Gallus gallus* genome in order not to lose elements. For instance the analysis of Signatures of regulatory activity in noncoding accelerated elements was performed using the chicken-specific regulatory data from the FAANG project (Pan et al. 2023) **available at (https://farm.cse.ucdavis.edu/~zhypan/Chicken_FAANG/Tissue_specific_regulator/) (see Results in pages 13 and 14).**

<u>Other comments</u>

Figure 1: mark up scale for the two trees either as time or as substitutions per site.

We have modified the trees included in Fig. 1, these trees are meant to show the species used and the relationships among them but they do not conserve the length of the branches.

Figure 2 c,d add unit on x-axis

Thanks for pointing out this mistake, in the revised version we have corrected it.

Reviewer #3 (Remarks to the Author)

Corrections suggested by this reviewer are highlighted in light blue in the revised manuscript.

In this manuscript, the authors perform a comparative analysis of different animal genomes looking for genomic regions that accumulate more changes than expected, according to the normal mutation rate. Using this methodology, they identify sets of highly accelerated regions specific for mammals and birds. They argue that the group of accelerated regions presented in this study for mammals is more accurate than the one previously published, thanks to the introduction of platypus in the genomic comparison in this work, while only therian species were used previously. Some of the accelerated regions identified are tested for regulatory activity using transgenic assays in zebrafish.

The authors claim that both, human versions of these putative regulatory regions and their orthologous sequence in alligator (taken as an outgroup of mammals) drive the reporter expression to similar domains, despite the good number of specific changes accumulated in mammals in these regions. They attribute this to compensatory changes in these sequences that maintain a functional conservation, although the sequence conservation is partially lost. This interpretation sounds too complicated to me, when the similarity among different founders' expression in some of the lines is controversial. The selected regions drive expression to very broad domains, including neural tissue, and this makes very difficult to evaluate whether the expression driven by two homologous regions (or even different insertions of the same region) are equivalent.

Furthermore, I find that this statement is not essential for the whole story, just probing that these regions act as enhancers is the important message, since it demonstrates their capacity to contribute to the emergence of specific innovations.

In addition, I think that these two messages are somehow contradictory: 1) accelerated regions are "critical for the evolution of specific traits"; and 2) mammalian and reptile orthologous regions drive similar expression patterns, which suggests that "some substitutions were probably fixed to compensate for other changes maintaining the expression pattern unaltered".

In my opinion, one of the most interesting conclusions of the manuscript is that accelerated regions have contributed to the emergence of specific novelties of each animal group. However, this conclusion is not supported by appropriate statistical analysis and figures, it is just based on the direct observation of GO terms enriched in genes associated to these genomic regions (for example: "This analysis showed us that those genes with clustering of ncMARs are more frequently related to typical mammalian traits such as lactation, diaphragm changes and dentition"). In order to support this and other conclusions, I would suggest thinking of a statistical analysis where the enrichment in GO terms specific for evolutionary novelties were compared to the enrichment in other terms related to ancestral traits, if such analysis is possible.

On the other hand, some unexpected GO terms are ignored, like ‘lactation behaviour’ in the shared accelerated regions between mammals and birds.

Besides, results from GO term enrichment analyses could be presented in some kind of friendly plots, like dotplots, in which some important values can also be shown (number of genes selected vs. genes in the category, pvalue, etc.).

Strengths of the work:
- The general approach and the methodologies used are correct and the conclusions of the study are interesting.
- Introducing a monotreme in the mammalian comparative is definitely useful in order to identify genomic characteristics present in the mammal last common ancestor.
- Performing functional assays help to validate the results, probing that accelerated regions have indeed regulatory activity.

Weaknesses of the work:
- The approach is not especially innovative.
- Language use is not good enough, it requires a strong review.
- Figures are repetitive and there are important conclusions that are not supported by them.
- Lack of statistical analyses that support some conclusions.

Taking all this into account, the manuscript should be strongly revised to deserve its publication in Nature Communication. As a summary, I will give a general outline of the main changes that should be made, in my opinion:
- Revision of language in general.
- Move some of the main figures regarding to transgenesis assays to supplementary info.
- Generate some new figure(s) with GO term enrichments as dotplots (or similar) and statistics.
- Perform a statistical comparative analysis to support GO-based conclusions, like the involvement of accelerated regions in the emergence of evolutionary novelties.
- Explain the contradiction of accelerated regions maintaining their function but contributing to novelties, or revisit some of these interpretations.

We thank the reviewer for its feedback. Below, we address each point raised and outline the steps we took to revise the manuscript:

1. **<u>Revision of language.</u> We appreciate your comments and we acknowledge the importance of improving readability of the manuscript. We have revised the language throughout the manuscript to ensure that the writing is clear and concise.**
2. **<u>Reorganization of figures.</u> To streamline the manuscript, we moved the figures including transgenesis results for the following accelerated regions (ncMAR1, ncMAR3, ncMAR4, ncMAR5, ncMAR7) to the supplementary information (now Supp. Figs. 3 to 7) and summarized the information in one main figure (now Fig. 4). We also summarized the *NPAS3* transgenesis assays in one main figure (Figure 5) and moved the two main figures containing HAR173/ncMAR506 and ncMAR3220/ncAvAR1232 results to Supp. Figs. 8 and 12, respectively.**

3. **<u>New Figures for GO Term Enrichments.</u> We generated dot plots, to present GO term enrichment results more effectively. These figures (Fig. 2 and Supp. Fig. 2) display more clearly the statistical significance and enrichment values of the GO terms in comparison to the background of conserved regions from which accelerated regions evolve. The new figures include statistics that aid interpretation. We additionally performed a simulation study and calculated empirical p-values to further support our conclusions (see below).**
4. **<u>Statistical comparative analysis to support GO conclusions.</u> Thank you for this insightful comment. We have modified the GO terms analysis in order to improve the comparison between conserved and accelerated regions. We understand the concern regarding the potential bias in GO analysis due to the non-random distribution of conserved elements around specific genes, as well as the risk of overrepresentation in certain GO categories. Following your suggestion in the new version we implemented a new strategy to analyze GO terms, comparing the results of rGREAT** (Gu and Hübschmann 2023) **obtained from the set of accelerated regions with those obtained from equivalent sets of conserved regions. To do that, first we sampled conserved regions, building 5,000 sets with the same size and chromosome distribution as the accelerated regions set, and evaluated rGREAT on each one of these sets. Then, we built an empirical distribution of 'n_hits' values for each specific go term from rGREAT results of each of the 5,000 conserved regions subsets. Next, we calculated the 'n_hits' value from the rGREAT result of the accelerated regions set (acc_observed_region_hits). Finally, we calculated the proportion of random n_hits observed that are greater than acc_observed_region_hits.**
If this proportion was less than 0.05, then this set of accelerated regions differs from the random sets of conserved regions evaluating this specific GO term, and we can assume that the behavior of the accelerated regions set is independent of its conserved nature for this term. We considered this proportion as an empirical p-value, and this p-value for each GO term was corrected for multiple comparisons. The Benjamini-Hochberg method was applied to compute approximate false discovery rates (FDRs) for each term.
We report a GO term as associated with the accelerated regions set if
a. The behavior of the accelerated regions set is independent of its conserved nature
and
b. The adjusted p-value returned by the binomial test performed by rGREAT on the accelerated regions set is less than 0.05.
We have updated the methods section <u>Gene ontology analysis</u> (pages 35 to 37) and we provide the code and more details of the implementation at (https://github.com/paulati/ncomms-24-47757/tree/main/go_analysis). The results of this new approach are included in Fig. 3, Supp. Fig. 2 and Supp. Tables 14-19, described in pages 11 to 13 of the Results section and further discussed in the Discussion section (pages 24 and 25). This new analysis shows that ncMARs are significantly enriched in genomic regions containing genes that play key roles in behavioral and cognitive processes (such as behavioral fear response, social behavior, learning or memory, suckling behavior), nervous system development (cerebral cortex regionalization or neuron migration), cell proliferation, differentiation, and migration (such as mammary gland epithelial

cell proliferation, endothelial cell migration, and skeletal muscle cell differentiation), and cardiovascular and muscular system development (including cardiac muscle tissue growth, regulation of heart growth, and ventricular cardiac muscle cell differentiation). Particularly interesting is the presence of several terms related to the morphogenesis of the heart, an organ that underwent intense remodeling in the mammalian lineage. In fact, we found ncMARs related to genes like *TBX5*, a key regulator of heart development (Koshiba-Takeuchi et al. 2009; Steimle and Moskowitz 2017) **and to other genes such as TBX2, TBX20, BMP2 largely involved in heart evolution in vertebrates** (Jensen et al. 2013)**. It is also noteworthy, the presence of terms related to brain development, particularly to cerebral cortex regionalization, linked to ncMARs associated with genes like *EMX1/2*** (Shinozaki et al. 2002) **and DMRTA** (Konno et al. 2012)**. We also found GO terms indicating significative enrichment in accelerated regions associated to genes involved in mammary gland development including *STAT6*** (Haricharan and Li 2014)**, *WNT5A*** (Roarty and Serra 2007)**, *GATA3*** (Slepicka, Somasundara, and Dos Santos 2021)**, etc.**
In addition, we found that ncAvARs were enriched in regions carrying genes related to nervous system development, immune response, regulation of hormonal and neurotransmitter pathways, cell proliferation and differentiation, intracellular components rearrangements and organ morphogenesis including middle ear morphogenesis, atrial cardiac muscle tissue morphogenesis and ventricular cardiac muscle cell development. It is interesting to note that genes related to the ear morphogenesis particularly the middle ear are significantly overrepresented in the ncAvARs group, particularly *EDNRA* (Ruest and Clouthier 2009)**, *EYA1*** (Zou et al. 2008) **and *NKX3-2*, a known transcription factor involved in the morphological evolution of the middle ear** (Anthwal, Joshi, and Tucker 2013)**. Among the terms related to nervous system development it is noteworthy the presence of cerebellar development genes linked to accelerated sequences including *LDB1*** (Zhao et al. 2007)**, *SKOR-2 (Nakatani et al. 2014)*, particularly because of the link between the cerebellum and the ability to power flight in birds** (Balanoff et al. 2024)**.**

5. **<u>Contradiction of accelerated regions maintaining their function but contributing to novelties.</u> We understand that you found these two points as contradictory and we appreciate that you pointed out this very important topic that allows us to better explain the message of our work.**
In fact we hypothesize that accelerated regions are critical for the evolution of specific traits through the modification of the regulatory machinery controlling the expression of key genes for the trait under study. Our functional analysis shows that mammalian and reptile orthologous sequences of four ncMARs drive similar expression patterns, which suggests that some substitutions were probably fixed to compensate for other changes maintaining the expression pattern unaltered. However, using this kind of comparative analysis as you pointed out, we can not rule out the possibility that changes in the cellular types or additional developmental stages where the enhancers are active exist and more detailed analysis will be necessary to assess these hypotheses. On the other hand, we observed that some of the functionally assayed ncMAR regions located in the *NPAS3* locus display differences in

expression when comparing the different ortholog sequences. In fact, we found that the human and chimpanzee ortholog of HAR173/ncMAR506 are active as transcriptional enhancers during development, whereas the alligator sequence is not active. These results suggest that changes in the mammalian lineage could have induced functional changes as a regulatory region. In addition, we noticed that the human sequence of ncMAR3220/ncAvAR1232 lost enhancer function since the reptile orthologs (chicken and alligator) are active as enhancers at several tissues across development, particularly in the nervous system. These results again suggest that the accelerated evolution process impacted on function in the mammalian lineage, although the sequence also underwent accelerated evolution in the bird lineage, also suggesting that acceleration does not immediately translate into functional changes. In summary, of the five *NPAS3*-ncMARs tested, two showed gain or loss of function in the mammalian lineage.

To summarize, our results indicate that accelerated evolution could have an impact in regulatory function but this is not always the rule. In our experience, between one third and half of HAR sequences display functional changes between human and chimpanzee orthologs when tested in enhancer assays in transgenic zebrafish or mice (Kamm et al. 2013; Berasain et al. 2024)**. Similar results have been also observed by other researchers who have found that 31 % of tested human and chimpanzee HAR orthologs show expression differences in transgenic mice** (Capra, Erwin, et al. 2013)**. In addition, this trend has been also observed using other experimental approaches like massive parallel reporter assays (MPRA) in different cell types. In fact, around 30 % of 1363 HARs and 50 % of 293 active HAR enhancers showed human-chimp differences when tested in neural stem cells** (Uebbing et al. 2021) **or in human or chimpanzee neural precursor cells (NPCs)** (Whalen et al. 2023)**.**
Thus, we can conclude that accelerated noncoding regions could potentially drive expression changes on target genes but functional studies are necessary to really assess this capability. This is a very important message that reinforces the idea that functional studies are keen to drive conclusions about the impact of sequence evolution on functional and morphological changes. In other words, bioinformatics analyses alone are not enough to fully understand the link between molecular and phenotypic evolution. We have broadened the discussion about this very important topic in the Discussion section on page 27.

References

Anthwal, Neal, Leena Joshi, and Abigail S. Tucker. 2013. “Evolution of the Mammalian Middle Ear and Jaw: Adaptations and Novel Structures.” *Journal of Anatomy* 222 (1): 147–60.

Balanoff, Amy, Elizabeth Ferrer, Lemise Saleh, Paul M. Gignac, M. Eugenia L. Gold, Jesús Marugán-Lobón, Mark Norell, et al. 2024. “Quantitative Functional Imaging of the Pigeon Brain: Implications for the Evolution of Avian Powered Flight.” *Proceedings. Biological Sciences* 291 (2015): 20232172.

Berasain, Lara, Paula Beati, Anabella P. Trigila, Marcelo Rubinstein, and Lucia F. Franchini. 2024. “Accelerated Evolution in the Human Lineage Led to Gain and Loss of Transcriptional Enhancers in the RBFOX1 Locus.” *Science Advances* 10 (26): eadl1049.

Bininda-Emonds, Olaf R. P., Marcel Cardillo, Kate E. Jones, Ross D. E. MacPhee, Robin M. D. Beck, Richard Grenyer, Samantha A. Price, Rutger A. Vos, John L. Gittleman, and Andy Purvis. 2007. “The Delayed Rise of Present-Day Mammals.” *Nature* 446 (7135): 507–12.

Boyd, J. Lomax, Stephanie L. Skove, Jeremy P. Rouanet, Louis-Jan Pilaz, Tristan Bepler, Raluca Gordân, Gregory A. Wray, and Debra L. Silver. 2015. “Human-Chimpanzee Differences in a FZD8 Enhancer Alter Cell-Cycle Dynamics in the Developing Neocortex.” *Current Biology: CB* 25 (6): 772–79.

Caporale, Alfredo Leandro, Alejandro R. Cinalli, Marcelo Rubinstein, and Lucía F. Franchini. 2024. “The Human Accelerated Region HAR202 Controls NPAS3 Expression in the Developing Forebrain Displaying Differential Enhancer Activity between Modern and Archaic Human Sequences.” *Molecular Biology and Evolution* 41 (10). https://doi.org/10.1093/molbev/msae186.

Caporale, Alfredo Leandro, Catalina M. Gonda, and Lucía Florencia Franchini. 2019. “Transcriptional Enhancers in the FOXP2 Locus Underwent Accelerated Evolution in the Human Lineage.” *Molecular Biology and Evolution*, July. https://doi.org/10.1093/molbev/msz173.

Capra, John A., Genevieve D. Erwin, Gabriel McKinsey, John L. R. Rubenstein, and Katherine S. Pollard. 2013. “Many Human Accelerated Regions Are Developmental Enhancers.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 368 (1632): 20130025.

Capra, John A., Melissa J. Hubisz, Dennis Kostka, Katherine S. Pollard, and Adam Siepel. 2013. “A Model-Based Analysis of GC-Biased Gene Conversion in the Human and Chimpanzee Genomes.” *PLoS Genetics* 9 (8): e1003684.

Christmas, Matthew J., Irene M. Kaplow, Diane P. Genereux, Michael X. Dong, Graham M. Hughes, Xue Li, Patrick F. Sullivan, et al. 2023. “Evolutionary Constraint and Innovation across Hundreds of Placental Mammals.” *Science (New York, N.Y.)* 380 (6643): eabn3943.

Feng, Shaohong, Josefin Stiller, Yuan Deng, Joel Armstrong, Qi Fang, Andrew Hart Reeve, Duo Xie, et al. 2020. “Dense Sampling of Bird Diversity Increases Power of Comparative Genomics.” *Nature* 587 (7833): 252–57.

Gu, Zuguang, and Daniel Hübschmann. 2023. “rGREAT: An R/bioconductor Package for Functional Enrichment on Genomic Regions.” *Bioinformatics (Oxford, England)* 39 (1). https://doi.org/10.1093/bioinformatics/btac745.

Haricharan, S., and Y. Li. 2014. “STAT Signaling in Mammary Gland Differentiation, Cell Survival and Tumorigenesis.” *Molecular and Cellular Endocrinology* 382 (1): 560–69.

Holloway, Alisha K., Benoit G. Bruneau, Tatyana Sukonnik, John L. Rubenstein, and Katherine S. Pollard. 2016. “Accelerated Evolution of Enhancer Hotspots in the Mammal Ancestor.” *Molecular Biology and Evolution* 33 (4): 1008–18.

Jensen, Bjarke, Tobias Wang, Vincent M. Christoffels, and Antoon F. M. Moorman. 2013. “Evolution and Development of the Building Plan of the Vertebrate Heart.” *Biochimica et Biophysica Acta* 1833 (4): 783–94.

Kaessmann, Henrik. 2010. “Origins, Evolution, and Phenotypic Impact of New Genes.” *Genome Research* 20 (10): 1313–26.

Kamm, Gretel B., Francisco Pisciottano, Rafi Kliger, and Lucía F. Franchini. 2013. “The Developmental Brain Gene NPAS3 Contains the Largest Number of Accelerated Regulatory Sequences in the Human Genome.” *Molecular Biology and Evolution* 30 (5): 1088–1102.

Konno, Daijiro, Misato Iwashita, Yoshiaki Satoh, Asuka Momiyama, Takaya Abe, Hiroshi Kiyonari, and Fumio Matsuzaki. 2012. “The Mammalian DM Domain Transcription Factor Dmrta2 Is Required for Early Embryonic Development of the Cerebral Cortex.” *PloS One* 7 (10): e46577.

Koshiba-Takeuchi, Kazuko, Alessandro D. Mori, Bogac L. Kaynak, Judith Cebra-Thomas, Tatyana Sukonnik, Romain O. Georges, Stephany Latham, et al. 2009. “Reptilian Heart Development and the Molecular Basis of Cardiac Chamber Evolution.” *Nature* 461

(7260): 95–98.
Liu, Anguo, Nini Wang, Guoxiang Xie, Yang Li, Xixi Yan, Xinmei Li, Zhenliang Zhu, et al. 2023. “GC-Biased Gene Conversion Drives Accelerated Evolution of Ultraconserved Elements in Mammalian and Avian Genomes.” *Genome Research* 33 (10): 1673–89.
Nakatani, Tomoya, Yasuko Minaki, Minoru Kumai, Chiemi Nitta, and Yuichi Ono. 2014. “The c-Ski Family Member and Transcriptional Regulator Corl2/Skor2 Promotes Early Differentiation of Cerebellar Purkinje Cells.” *Developmental Biology* 388 (1): 68–80.
Pan, Zhangyuan, Ying Wang, Mingshan Wang, Yuzhe Wang, Xiaoning Zhu, Shenwen Gu, Conghao Zhong, et al. 2023. “An Atlas of Regulatory Elements in Chicken: A Resource for Chicken Genetics and Genomics.” *Science Advances* 9 (18): eade1204.
Roarty, Kevin, and Rosa Serra. 2007. “Wnt5a Is Required for Proper Mammary Gland Development and TGF-Beta-Mediated Inhibition of Ductal Growth.” *Development (Cambridge, England)* 134 (21): 3929–39.
Ruest, Louis-Bruno, and David E. Clouthier. 2009. “Elucidating Timing and Function of Endothelin-A Receptor Signaling during Craniofacial Development Using Neural Crest Cell-Specific Gene Deletion and Receptor Antagonism.” *Developmental Biology* 328 (1): 94–108.
Seki, Ryohei, Cai Li, Qi Fang, Shinichi Hayashi, Shiro Egawa, Jiang Hu, Luohao Xu, et al. 2017. “Functional Roles of Aves Class-Specific Cis-Regulatory Elements on Macroevolution of Bird-Specific Features.” *Nature Communications* 8 (February):14229.
Shinozaki, Koji, Toshihiko Miyagi, Michio Yoshida, Takaki Miyata, Masaharu Ogawa, Shinichi Aizawa, and Yoko Suda. 2002. “Absence of Cajal-Retzius Cells and Subplate Neurons Associated with Defects of Tangential Cell Migration from Ganglionic Eminence in Emx1/2 Double Mutant Cerebral Cortex.” *Development (Cambridge, England)* 129 (14): 3479–92.
Slepicka, Priscila Ferreira, Amritha Varshini Hanasoge Somasundara, and Camila O. Dos Santos. 2021. “The Molecular Basis of Mammary Gland Development and Epithelial Differentiation.” *Seminars in Cell & Developmental Biology* 114 (June):93–112.
Steimle, J. D., and I. P. Moskowitz. 2017. “TBX5: A Key Regulator of Heart Development.” *Current Topics in Developmental Biology* 122:195–221.
Stiller, Josefin, Shaohong Feng, Al-Aabid Chowdhury, Iker Rivas-González, David A. Duchêne, Qi Fang, Yuan Deng, et al. 2024. “Complexity of Avian Evolution Revealed by Family-Level Genomes.” *Nature* 629 (8013): 851–60.
Uebbing, Severin, Jake Gockley, Steven K. Reilly, Acadia A. Kocher, Evan Geller, Neeru Gandotra, Curt Scharfe, Justin Cotney, and James P. Noonan. 2021. “Massively Parallel Discovery of Human-Specific Substitutions That Alter Enhancer Activity.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (2). https://doi.org/10.1073/pnas.2007049118.
Whalen, Sean, Fumitaka Inoue, Hane Ryu, Tyler Fair, Eirene Markenscoff-Papadimitriou, Kathleen Keough, Martin Kircher, et al. 2023. “Machine Learning Dissection of Human Accelerated Regions in Primate Neurodevelopment.” *Neuron* 111 (6): 857–73.e8.
Whalen, Sean, and Katherine S. Pollard. 2022. “Enhancer Function and Evolutionary Roles of Human Accelerated Regions.” *Annual Review of Genetics* 56 (November):423–39.
Zhang, Guojie, Cai Li, Qiye Li, Bo Li, Denis M. Larkin, Chul Lee, Jay F. Storz, et al. 2014. “Comparative Genomics Reveals Insights into Avian Genome Evolution and Adaptation.” *Science* 346 (6215): 1311–20.
Zhang, Guojie, Carsten Rahbek, Gary R. Graves, Fumin Lei, Erich D. Jarvis, and M. Thomas P. Gilbert. 2015. “Genomics: Bird Sequencing Project Takes off.” *Nature* 522 (7554): 34.
Zhao, Yangu, Kin-Ming Kwan, Christina M. Mailloux, Woon-Kyu Lee, Alexander Grinberg, Wolfgang Wurst, Richard R. Behringer, and Heiner Westphal. 2007. “LIM-Homeodomain Proteins Lhx1 and Lhx5, and Their Cofactor Ldb1, Control Purkinje Cell Differentiation in the Developing Cerebellum.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (32): 13182–86.
Zou, Dan, Christopher Erickson, Eun-Hee Kim, Dongzhu Jin, Bernd Fritzsch, and Pin-Xian Xu. 2008. “Eya1 Gene Dosage Critically Affects the Development of Sensory Epithelia in

the Mammalian Inner Ear.” *Human Molecular Genetics* 17 (21): 3340–56.

RESPONSE TO REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

Review of Comparative genomics sheds light on mammalian and avian gene regulation and phenotypic evolution

Comment (C)1: First of all I would like to acknowledge the extensive work done by the authors to revise the manuscript based on the three reviewer comments with detailed response to each point raised. This has definitely improved the manuscript a lot.

Response (R)1: Thanks to the Reviewer for his comments and suggestions that have greatly helped to improve our paper.

C2: Fig1:
Upon going through the manuscript again, one thing that caught my eye which I did not see on the first time around is the bird tree in Fig 1. That tree is not the correct phylogenetic relationships of birds. The first branch is correct with Struthio and Tinamus. But then according to this tree, the next branch is Picoides, a woodpecker. Then there are two clades, one containing a hodgepodge of random birds and a second that in some respects looks like the correct coarse relationships among birds but with some glaring issues as well. The woodpecker is not the earliest branching Neoaves. There is no support for this two group branching in literature. The galloanserae (chicken and geese) are embedded somewhere within group 1 and broken up and the galliformes are sister to an owl (which itself is separated from other birds of prey in clade 2). Also, in clade 2, the falcons are sister to birds of prey (not to parrots). The rifleman is not the earliest branch of the passerine birds (instead it is the ground jay). Where did you get this tree? How did this tree happen?
Second, minor point, the branch labelled mammals on the figure is in the wrong spot (should be one more node down, currently the highlighted branch is the reptile branch).

R2: Thanks to the reviewer for noticing this mistake in Fig.1. Since the aim of the tree included in this figure is to display the species used and also to indicate the branch tested, the right tree to include is the consensus species tree. We have done it now, following the tree published by the Bird 10K Project (Stiller et al. 2024)**.**
The original tree we displayed was built from the 77-way alignment. This tree model was generated using the phyloFit program from the PHAST package (REV model, EM algorithm, medium precision) using multiple alignments of 4-fold degenerate sites extracted from the 77-way alignment. The aim of this tree in the method we used to detect accelerated sequences is to indicate the branch or branches to be tested. In our case since the branch tested is the one of the split between the outgroup (no birds vertebrates) and Birds, the tree used does not affect the results since this branch is well defined.
Thanks also for noticing the mislabeling of the mammalian branch, which is now corrected in the revised Fig. 1.

C3: Methods: With regards to the methodology, thank you for clarifying your methods more and providing a more detailed description of the code and methods. It makes more sense

now as to how you are defining acceleration. Also, yes the paucity of alignments containing all amniotes warrants the use of the 100 way and 77 way alignments and even though these alignments may have issues due to the quality of samples/genomes, it hopefully does not affect conserved regions as much.

R3: Thank you for your comments that led us to display our methodology in a more clear way.

C4: Minor point: in the methods the common names of species are inconsistently written down with uppercase and lowercase names (e.g. Hooded Crow vs american eagle). I would advise just standardizing the names based on the journal recommendations.

R4: Thanks for this comment, we have corrected all common names to be accurate and consistent all throughout the manuscript. The section containing the corrected names is highlighted in purple in the edited version of the manuscript (pages 29 to 31).

C5: Also, the gBGC analysis helps clarify some of these accelerated elements. Thank you for taking the time to address that.

R5: Thanks for your suggestion, because it improved our paper a lot.

C6: Results:

Line 127: I would change the word basal with earliest diverging. The word 'basal' has been controversial since these taxa are still currently living and not 'basal' in the traditional sense of the word. You can have a basal node (the node where the paleognaths and neognaths split) but not a basal bird (unless you literally mean the bird that lived 170 million years ago).

R6: Thanks for noticing this issue, we have edited replacing the word "basal" with "earliest diverging" following your suggestion (page 5, highlighted in purple in the last submitted version).

C7: On a related note. I suspect you dropped from 155,630 elements to 5,659 elements in the bird dataset because the tinamous seems to have a highly accelerated rate of evolution, the often appear in this long branch on the avian tree and so many more fixed differences are probably in the tinamous lineage. My guess is that if you made your criteria as only applying to ostrich, you'd find more shared conserved elements that meet your criteria. This also explain the discrepancy you address for reviewer 2. Just worth a check.

R7: We acknowledge your insightful comment. In our approach to detect Aves accelerated elements, we required that at least one early diverging bird from Palaeognathae (either white throated tinamou, *Tinamus guttatus* or ostrich, *Struthio camelus*) be present in all alignments. In addition, in our method the tinamou is used to classify accelerated sequences (i.e. nucleotide changes detected comparing tinamou to the background block of species (no bird vertebrates) must be conserved in the foreground block of species (Birds) to be considered as accelerated changes). We choose tinamou as a classifier because it is considered one of the earliest diverging bird species (Almeida et al. 2022; Bertelli, Chiappe, and Mayr 2014) **and also because its genome has higher coverage than that of the ostrich. If tinamou has an accelerated rate of evolution, probably many changes detected in tinamou would not be conserved in the block of other birds. That probably translates in a lesser amount of accelerated sequences detected and in that sense probably our method is conservative. We consider your suggestion of using ostrich as a classifier very appealing, but changing this criteria in the current paper will probably modify some of**

our results and because of that we prefer to maintain tinamou as our criterion for classification.

References

Almeida, Francisca C., Ana L. Porzecanski, Joel L. Cracraft, and Sara Bertelli. 2022. “The Evolution of Tinamous (Palaeognathae: Tinamidae) in Light of Molecular and Combined Analyses.” *Zoological Journal of the Linnean Society* 195 (1): 106–24.

Bertelli, Sara, Luis M. Chiappe, and Gerald Mayr. 2014. “Phylogenetic Interrelationships of Living and Extinct Tinamidae, Volant Palaeognathous Birds from the New World.” *Zoological Journal of the Linnean Society* 172 (1): 145–84.

Stiller, Josefin, Shaohong Feng, Al-Aabid Chowdhury, Iker Rivas-González, David A. Duchêne, Qi Fang, Yuan Deng, et al. 2024. “Complexity of Avian Evolution Revealed by Family-Level Genomes.” *Nature* 629 (8013): 851–60.