nature portfolio

Peer Review File

Drivers of avian genomic change revealed by evolutionary rate decomposition

Corresponding Author: Dr David Duchene

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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
Review for: Drivers of avian genomic change revealed by evolutionary rate decomposition

The presented work by Duchêne et al is a fascinating exploration of recently available whole genome data for birds. In general, I really enjoyed reading this paper, and I think it will interest a broad audience. It is exciting that we are finally moving into an era in which this level of detailed analysis can be attempted at such a wide taxonomic scale, further extending our reach into the past. There are many detailed and nuanced analyses reported here, so it will take some time and dedication for any reader to fully appreciate them. It is possible that some of my comments are a consequence of details I missed; I apologize to the authors in advance if so.

I have read through the paper several times, and I think I have been able to follow the primary narrative. The paper has two parallel analytical tracks; the first is to examine how a suite of traits is related to different aspects of molecular rate variation. I was very impressed by the evaluation of 29 traits, which is indeed comprehensive. The second track is to “decompose” different aspects of molecular rate variation across individual genes and phylogeny branches to assess which, if any, have “an unusually large impact on rate variation.” The authors employ a technique implemented in the software ClockStarRX, developed by the authors. This technique is available, though it currently exists as an unreviewed bioRxiv preprint.

I have some thoughts, comments, and concerns on both tracks, so I will split my review into two sections. Most of my comments are related to the methods associating traits with molecular rates. Then, I will provide some more minor comments. While this paper and overall approach have tremendous merit, addressing the issues I highlight below will significantly strengthen the paper’s conclusions.

Linking trait variation to individual patterns of molecular rate variation:

The authors’ approach is to first examine each of a set of 29 traits individually with respect to each aspect of molecular evolution (dN, dS and w), using “simple” phylogenetic regression.

1) I assume this means that the regression model used assumes residual variation follows a Brownian motion process (I believe this is the default for phylolm)? Not clear what model is used for these regressions. If so, I suggest the authors specify the methods, as it is a key assumption.

2) Figure 2 indicates several variables that are “mass corrected,” but no details about this are provided in the methods. Why is mass-correction important in this context? Why was clutch size not mass-corrected if the other highlighted variables were? This is an important point to clarify because clutch size is highlighted as the most important predictor of molecular rate variation. Revell 2009 (‘Size-correction and principal components for interspecific comparative studies’) provides useful guidance.

2a) A lesser point is that clutch size is probably not optimally modeled as a continuous trait (it is a type of count data). Although many authors do this, it is better modeled with a Poisson or binomial regression for count data. These can be implemented using the BRMS R package while accounting for phylogenetic signals. In most situations, approximating clutch size as a log-transformed continuous trait is probably fine, but I emphasize this here because clutch size is noted as the most

significant feature in the set of traits; thus, this part of the paper hinges on the validity of this result.

3) I am concerned about the statistical validity of taking the variables that are identified as being significant in 1:1 regression analysis (e.g., clutch size vs w) and then using them together in multiple regression. The intention of this latter aggregation in multiple regression is not clearly articulated; the authors say, lines 487-490: "We then ran multiple phylogenetic regression, where all the significant explanatory traits from the individual regressions were added as explanatory variables as partial terms, testing whether significant variables explained molecular rates above the effect of other traits." I think this is saying that the tests are designed to evaluate whether a focal dependent variable (e.g., clutch size) remains significantly associated with the independent variable (e.g., w) after controlling for the effects of other variables identified in 1:1 regression. If this is the correct interpretation, it could be more clearly stated. This may be a fine thing to do, but as no examples of model comparison are provided (e.g., comparing simpler to more complex models through their likelihood or AIC), it is challenging to assess the justification for running the multiple regression models. In general, it seems preferable to include multiple predictors when their inclusion is justified in terms of relative model fit to the data or when there are stated hypotheses about multiple or interaction effects.

Moreover, it is unlikely that these traits evolve independently; they will be correlated and will follow a complex multivariate variance-covariance structure. I suspect the authors know this, as noted above, there are some variables that are qualified as being “mass corrected” (though this is not explained). Multivariate phylogenetic regression techniques are available in R packages like mvMORPH (see mvgls) or SLOUCH, for example, and could be employed here to assess how the multivariate relationships among these traits are related to various evolutionary rates. I recognize this is a somewhat distinct research area and do not ask the authors to explore it (unless they want to); only to better acknowledge that focusing on individual traits may be misleading when traits covary in complex ways.

4) I was surprised by the result that body mass is apparently not related to any metrics of molecular evolution, as this has been reported in numerous other studies, including those which employ genome-scale data (e.g., Botero-Castro et al 2019; 'Avian Genomes Revisited: Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds'). [e.g., Lines 149-151, I assume this means all types of molecular rates and not only w? Apologies if I have misunderstood]. It could be useful for the authors to speculate further on what might explain this result. This is especially surprising in the context of 1:1 regression analyses that evaluated body mass and did not control for covariance with other traits (which I would typically guess would remove the effect of body mass). This seems somewhat in conflict with results reported in the associated paper (Stiller et al.), which report an increase in substitution rates associated with the end-Cretaceous transition and an associated decrease in body mass. The paper reports (Fig 2b) that “log(mass-corrected generation length)” is associated with substitution rates, but not if “log(generation length)” is associated (hence the importance of clarifying the issues of mass correction noted above). log(body mass) correlates well with log(generation length) in birds, and since generation length is detected as a significant predictor, this pattern is indeed perplexing. It is also somewhat paradoxical, as body mass is expected to be a reasonable proxy of effective population size through the relationship body size has with environmental carrying capacity. It would be useful for the authors to check some simple metrics of this expected association, for example, checking if log(root to tip branch length) [from the gene trees] is related to log(body mass) (with both phylogenetic and non-phylogenetic regression). Perhaps another option is to be more circumspect about this result in light of other papers that use more sophisticated modeling approaches to detect associations (below).

5) The finding for beak depth is fascinating; it would be helpful for the authors to briefly clarify what this feature is (in context), as it may not be apparent to non-bird experts. Lines 128-131: I am surprised by the inclusion of hummingbirds on this list, as I generally think of them as having relatively long, high aspect ratio beaks relative to their body size, but maybe this is just me misunderstanding what ‘beak depth’ means. I suspect this makes more sense if we are talking about beak depth relative to body size, which seems to be the focus (in Fig 2). Could beak depth be a proxy for body size (if not mass)?

6) In general, although the huge scale of the work justifies simplifying analytical assumptions, the authors should probably acknowledge as a caveat somewhere that the ‘simplified’ regression approach makes many assumptions that can lead to biased inference. Although it would be impossible to use on a dataset of this scale, alternative multivariate approaches should be much less biased and more sensitive. One example is the model implemented in the software ‘coevol’ (Lartillot and Pujol 2011); this could be used to spot-check some results.

Linking trait and lineage variation to decomposed aspects of principal components:

1) I struggled a bit more with these sections, as they make use of an analytical pipeline employed in the ClockStaRX software, currently reported in a bioRxiv preprint (https://www.biorxiv.org/content/10.1101/2023.02.02.526226v1.full). The inclusion of Extended Data Figure 1 was very helpful to understand exactly what is happening here. It appears that there is some preliminary overlap between analyses presented in the preprint and those reported in the current manuscript, as the preprint investigates some similar patterns in avian phylogenomic data. In the main text section introducing these results [Line 184], I think it would be helpful for the authors to insert one additional short sentence providing more context for what this analysis is doing. The current language was not sufficient for me on my initial read, and I had to go to Extended Data Fig 1 to understand that the input data matrix into the PCA is a matrix of branches (columns) and genes (rows). I inferred that these analyses were performed independently for dN, dS, and w? It is not clearly stated if these parameters were analyzed jointly or independently. Further, can the authors clarify if the underlying PCA in ClockStaRX is a Phylogenetic PCA? I am not sure if it is necessary in this context, but as these data are all inherently comparative, it may be appropriate.

2) In general, I found these results to be interesting, and compelling and they convinced me that this approach might be a good way forward for tackling some of the gene-phenotype associations observed at this broad scale. While most

conclusions seem sound and well justified, I am concerned about one important methodological choice and its impact on some of the major results. The authors note in the methods that all gene trees were estimated on a fixed topolgoy indicative of the species tree (thus, they only vary in branch lengths). This was done so that more data could be evaluated for deeper nodes, as fewer and fewer concordant edges will exist as we go back in time. While I appreciate this fact, the downside of this choice is that it may introduce bias, resulting in some of the exact patterns that the paper reports (e.g., fast rates of molecular evolution associated with the origins of ancient clades, exactly where we expect there to be a lot of gene-tree discordance). The authors cite Mendes and Hahn (39), so they are aware of this potential source of bias. I personally believe that these patterns are likely real, but because the authors do not directly account for gene-tree topology discordance here, these results could be an artifact of incorrectly mapping substitutions to edges that do not exist in the true gene trees. As a solution, I wonder if it would be particularly challenging for the authors to provide as supplementary data a small vignette corroborating one or two of the interesting edges that are detected (e.g., the stem branches of Telluraves or the new Elementaves) with analyses that allow the gene tree topologies to vary? This would help give confidence that these results are not artifacts or biased in the vein of Mendes and Hahn.

3) I loved the discussion of the chromosome-scale variation, the importance of microchromosomes, their gene content, and the recombination landscape. Great stuff: this is really important!

Minor comments/questions by line:

Line 53:
The authors may wish to acknowledge body mass (Weber et al. 2014, Botero-Castro et al. 2019 (body mass), Berv and Field 2018 (body mass, metabolic rate) (cited elsewhere in the manuscript)), even though their current results do not support it.

Line 64-65:
Accelerated rates could be indicative of relaxed purifying selection in coding regions; as written, this phrase sort of implies the opposite.

Line 71:
“Modelling” is spelled with two l’s — there are a few other spots where this is done. Not sure if it is a typo or British spelling.

Line 102:
Generation length is correlated with which type of molecular rate? Not clear.

Line 134-35:
Were any other taxonomic exclusions made based on extreme values?

Line 158:
Can the authors clarify, w is not associated with any traits in the dataset?

Line 164-71:
The proposed explanation with respect to population size does not seem consistent with results reported in the other submitted paper (Stiller et al.), which report lineages associated with population size expansion associated with the end-Cretaceous extinction. Surely, the impact of Ne varies across the genome, but it's not clear how this might impact the overall proposed associations with body size.

Line 210-214:
How can this be true if there is no association between body size and variation in molecular rates? Seems like a paradoxical argument.

Line 215:
The authors may want to cite Brown et al. 2008 (Strong mitochondrial DNA support for a Cretaceous origin of modern avian lineages) and Berv and Field 2018 (Genomic Signature of an Avian Lilliput Effect across the K-Pg Extinction), who also investigated this idea. Berv and Field 2018 is cited elsewhere, so it would not increase the reference count to acknowledge that idea here.

Fig 2 caption; line 220:
It would be helpful to be specific about “locus rates” — e.g., that this means dN, dS, and w (for readers who focus on the figures).

Line 272:
The phrase “under high mutational pressure synteny” - I think there is a missing comma here: “under high mutational pressure, synteny”?

Line 290:
The reported relationship between w and tarsus length is pretty tenuous; the effect size is low (R2=0.064). With this type of regression modeling framework and the huge sample sizes used here, it is not very difficult to get a “significant” result. It may be helpful to know if including tarsus in a model predicting w significantly improves model fit (through something like likelihood or AIC).

Line 306-307:
As noted above, I am concerned that these observations of multiple early bursts are artifacts of the choice to fix the gene tree topologies to the species tree topology (even though I think they very well could be real).

Line 308:
'Accelerated rates' - is this referring to w?

Line 310-313: very cool result — I believe patterns of development/growth could be implicated along these deep branches and these results support that idea.

Line 320:
"signalling" see above; possible typo

Line 476-483:
Please be explicit about any data transformations, scaling, etc.

Line 484:
(a less minor point)
Using the family-level averages for phylogenetic comparative data/analyses is probably not a good practice, as this can bias estimates of the evolutionary parameters estimated in the phylogenetic regression models. E.g., if the family level average is not close to the family level MRCA (which it very well may not be, depending on clade age and phylogenetic structure within families), the model may have to explain trait variation for a given terminal, depending on the direction of the bias. In general, in this situation, it would be optimal to use the actual tip-level species data and then resample different species data from within each family to see if the results are robust (though, I would advise the authors use actual tip-level species data and not to bother with resampling, as this would already be an accurate representation of the terminal trait data with respect to the genomes underlying those terminals). This is particularly important here as there is some attempt to associate these traits with underlying patterns of genetic variation. As is, it is difficult to know what kinds of bias this may have introduced.

Line 511-512:
Related to my prior comment. Although the terminals here are sampled to represent avian families, it may be preferable to describe this as loadings of species. But, given the way the trait data are averaged, I am not sure.

Referee #2

(Remarks to the Author)
A) The manuscript "Drivers of avian genomic change revealed by evolutionary rate decomposition" attempts to identify life-history traits that correlate with rates of genomic changes in avian species. Specifically they look at protein evolutionary rates as well as intergenic regions. For their large scale analysis the authors rely on two datasets, one from Feng et al (2020) for the alignment data and Stiller et al (unpublished) for the tree data. The latter paper has been attached as supporting information and as far as I understand has been submitted else well. The authors complement these two datasets with trait data which they have collected from databases. The information for this are included in a spreadsheet available in a repository, and many of them rely on Tobias et al, 2022. Please note that the Feng data uses 1Kb pieces of intergenic sequences rather than whole genome alignments. While the underlying data has been used from other sources, the main innovation in this work is a decomposition method that would incorporate gene trees and species trees. They then use permutation tests to assess statistical significance.

B) Generally I think this is an interesting piece of work. It uses a large amount of data and attempts to identify general drivers of molecular evolution in avian species.

What I particularly like:
- Extended Figure 1 explaining their approach
- The speculation on tarsus length as an indicator of how panmictic/moveable individuals of a species are

What I think is missing:
- more focus on non-coding elements
- including variation into the genetic analysis (such as heterozygosity)

C/D/F) What I think is important to consider and which I could not find info on

1) How do you resolve when a gene tree and species tree disagree?
2) GC content as an important variable is considered in the text, however it is not really tested for. Principally this could easily be incorporated into the study. So for example how do the top 20% of genes (e.g. Fig4) compare regarding their GC content relative to the rest of the genes.
3) I think DNA methylation driven mutation rates get little attention. If GC content is high, so will be the CpG content which leads to increased mutation rates.

4) It has often been noted that GC content might not be at equilibrium in birds. This has drastic consequences if not incorporated into the models of evolution which usually assume a stationary base composition. In particular for long branches and synonymous sites this could be an issue. Codeml does have models to test for this.
5) For fig 3 - what is the proportion of genes that are in the same chromosomal location versus those that are not (e.g. due to intrachromosomal rearrangements)
6) How did you deal with chromosome number
7) Is there any analysis that considered ends of chromosomes, usually ends of macrochromsomes are very similar to microchromosomes
8) Fo fig 4 you identify many RNA related genes (ribosomes, splices

General comments:

- I think it would be good if the number of sites analyzed could be considered. For example in Fig 3 you mention number of coding and n intergenic. Is this the number of codons/genes/blocks - how many actual sites where included?
- What has been done if there is variation in the chromosome number, such as extra chromosomes - in general karyotypes are conserved, but some variation is still present (e.g. falcons)
- How did you proceed with chromosome 1 (I assume you used a pseudo 1 for songbirds)
- There is no mentioning of the GRC and potential duplication outsourcing of genes in species that possess it?
- where non coding RNAs considered at any point?
- Did you have any look at TE elements as genomic predictors (as these stopped spreading at some point)
- Have you tried thinning the data? Your point estimates for each branch will get worse if you add more species. I understand you used a lower threshold for excluding genomic elements - but I think fewer species will perhaps allow you more meaningful fun omega ratios

Line 375 form = from

Referee #3

(Remarks to the Author)
Summary of the key results:

This study utilises available genomic data, and morphological, ecological and life-history data, from 281 species of birds representing 87% of avian families. The genetic data come from the B10K community (Fang et al.) and the species attributes from the literature. The genomes are aligned and sequence variation between lineages is estimated. Some lineages have higher substitution rate than others; the question is why? Is it possible to find drivers/predictors for the variation in substitution rate among the selected species data? Evaluating a large number of morphological, ecological and life-history traits increases the chance of finding associations. The study lacks hypotheses, which may be considered a problem in this kind of parameter rich correlative studies as any result may be open for quite speculative interpretations. Several correlations between the four genetic metrics (dS, dN, omega, inter-genetic) and the species and chromosomal attributes are detected (including with clutch size, beak size and chromosome type).

Originality and significance:

To my knowledge these data sets have not been combined and analysed in this way before.

*Data & methodology: validity of approach, quality of data, quality of presentation;
Appropriate use of statistics and treatment of uncertainties;
Suggested improvements: experiments, data for possible revision;
References: appropriate credit to previous work?:*

Genetic data are from B10K (Fang et al.). This data set includes ≈370 species. Why are only 281 included here? Why were almost 100 species excluded? The authors must make a transparent statement of which were excluded and why. Maybe I missed this information?

The genetic variation is measured as the rate of synonymous dS and non-synonymous dN in genic regions, and degree of substitutions in inter-genic regions. Also, the ratio between dN and dS, termed omega, is utilised – this measurement is supposed to be the most appropriate estimate of selection as other evolutionary processes such as genetic drift and mutation rate are expected to affect dN and dS in the same way. Note, however, that variation of the degree of selection and drift in different time periods might cause much more complex patterns than explained by these simple metrics at a specific time point. Overall, omega is the most appropriate metric to measure putative drivers of substitution rates in the genome unrelated to genetic drift and mutation rate. The section outlining the substitution rates (lines 88-95) contains errors I believe and needs to be revised and references given. “We tested the association between these traits and average family-level rates across four measures of molecular evolution: dN, dS, dN/dS (ω), and rates in intergenic (IG) regions. These measures have the potential to provide complementary insights into the evolutionary process: dN is influenced by the mutation rate, population size, and selection; dS is influenced by the mutation rate; and their ratio (ω) is influenced by selection and population size. Genome-wide variation in ω is expected to be driven by population size. Rates in intergenic regions of the genome are expected to reflect the mutation rate.” Kimura’s original derivations for dN and dS both contain 1/N, and thus N is not supposed to affect omega (i.e., contrary to the suggestion above). This probably affects many of the interpretations in

this MS. Also rates in intergenic regions are affected by N.

The paragraph regarding the (lack of) results for omega is surprisingly vague. Basically, the logic and implications for most of the text at lines 158-171 are difficult to follow: “The lack of support for any association of genome-wide ω with body size or any other trait points to a limited role of variations in population sizes on avian molecular evolution. Population sizes are believed to have increased rapidly immediately after the K–Pg transition, during a period of newly available ecological niches. If this were the case, then we might also expect lower ω in smaller-sized avian taxa, consistent with the Lilliput effect, due to the more efficient removal of slightly deleterious nonsynonymous mutations24. One explanation for the lack of this signal in our data is that the effects of population-size fluctuations can be seen primarily on shallower timescales rather than in deep lineages where the signal has eroded. Alternatively, our finding of variation in ω across axes of variation and across chromosomes suggests that the impact of population size might be uneven across the genome (Stiller et al. submitted). Lastly, incomplete lineage sorting and reticulate evolution can also degrade the signal of ancestral population size39. Future examinations of these questions at finer taxonomic levels will be able to show a more nuanced picture, for instance by extending estimates of population size dynamics40 into deeper timescales.” Why is the “K–Pg transition” interesting here? Why are “smaller sized taxa” discussed? What is the “Lilliput effect”? Why is “selection” suddenly discussed, and would not omega have been associated with some traits if selection was more prominent? What does “variation in ω across axes of variation” mean; variation explains variation? What is and why are “incomplete lineage sorting and reticulate evolution” discussed? I believe that the degree of incomplete lineage sorting is probably quite limited at the family level (most branches are >15 Myrs).

Line 195 and following lines. I suggest discussing the actual level of dN and omega as well, as low and high values are associated with very different processes. The statement at line 205-206 needs a reference. Overall, you need to interpret PCs with caution (especially the ones explaining little of the variance). Line 208: To understand lineage specific patterns, it would be important to evaluate whether the multi-species alignment underlying these analyses is equally robust for different parts of the phylogeny. You need to discuss and clarify for the reader whether or not such potential issues might have affected your estimates of substitution rates.

Figure 3. I’m not sure this figure shows “mean chromosome rate”? (3A) I assume positive values mean higher than estimated with permutation. It is surprising to see dN>dS, and positive omega, for a large chromosome like 7. Also, this is the only chromosome with a large omega value, right? Could this be due to alignment errors, or translocations etc., in some part of the phylogeny. Note that chromosomal synteny is not conserved among all birds (there is variation in chromosome numbers, there are fused and split chromosomes, etc.). (3B) I think it is difficult to understand what all these PCs are telling (16 PCs (!) for intergenic are presented; e.g. what does PC 11 tell us?).

Line 257-260. “This finding on ω rates is consistent with widespread selective constraints on microchromosomes immediately after the K–Pg transition. This result is also consistent with the critical role that has been proposed for microchromosomes in harbouring housekeeping genes.” You need to explain how these conclusions arise, and how they are compatible with each other and previous results and statements for the K–Pg transition (this is now unclear).

Lines 263-265. “The overrepresentation of microchromosomes in the evolutionary rate shifts along early neoavian branches suggests that they have had a greater influence on avian evolution than macrochromosomes.” This is a very simplified conclusion and interpretation (possibly based only on visual screening?) of the results. Note that, in Figure 3A, chromosome 7 is the only chromosome with deviating omega – not the microchromosomes. In figure 3B some but not all microchromosomes show deviation signals for omega, and chromosome Z is also deviating, and only for some (few) PCs. Some of the chromosomes of intermediate size are showing similar patterns as the microchromosomes.

Lines 271-274. “Since our results of molecular rates focus on evolution at the nucleotide level, they suggest that even under high mutational pressure synteny can be conserved across deep timescales in microchromosomes, reflecting a balance between high nucleotide mutation rates even under severe selective constraints.” What is “they” referring to? “…high mutational pressure synteny can be conserved…” I think you have the data to find support this statement (or not) in your alignment. You need to make it clear why high mutational rate is suggested (omega tells primarily about selection, as mutation rate is thought to affect dS and dN similarly).

Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions; Conclusions: robustness, validity, reliability:

Line 276 and forward. How was “the most prominent branches driving rate variation” defined? Also, there are so many options for gene enrichment analysis in this data set – the four variables (ds,… etc) and PCs (up to 16 for intergenic). This makes the results of this whole exercise quite difficult to evaluate. Any genome-wide genetic signal will show some genes being more associated with the signal than other genes – thus, enriched pathways can always be presented and suggested, and are not particularly strong evidence for processes in correlative studies, although it may look like there is a robust molecular basis (as e.g. implicated by the results presented in figure 4, and in the Abstract).

A potential main concern with this paper is the use of deep lineage genetic data and shallow data for the morphological, ecological and life-history data. Or in fact it is not clear when reading the main body of the MS whether the morphological, ecological and life-history data are from the species between included in the analyses or whether these data are reflecting properties of the clades? This would have been more appropriate as the genetic data (on the family level) likely reflect processes taking place long before the actual species appeared. This might be why the authors find most (putative) signs of selection at the terminal branches (which are in fact representing several million years of evolution; probably in most cases

>15 Myrs).

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
Review for: Drivers of avian genomic change revealed by evolutionary rate decomposition

This is the second time I have reviewed this paper, and while it has improved in some respects, I still have concerns that prevent me from recommending publication in its current form.

Below, I provide some responses to the responses that the authors have offered to my original review, and then I follow with additional comments/questions going line by line.

That said, the most significant concern I have with this manuscript is that many of the regression results emphasized by the authors are “statistically significant” but with low effect sizes (my guess; see my line-by-line comments below). I have not been able to figure out where to access ‘Supplementary File 3,’ (Tables of phylogenetic regression data and results), so it's not clear if further information reported there would help me evaluate these aspects of the work.

In regression analyses, arbitrarily low p-values can be achieved (even considering correction for false discovery rate) with a dataset of this magnitude simply due to extreme sensitivity (very small effects can be detected). On one hand, this is a potential benefit of a dataset like that employed here. However, on the other hand, if many of the results represent minor effects (even if statistically significant), the significance of the overall conclusions is weakened.

For example, I am concerned by the lack of reporting of effect sizes for every test I can see in this manuscript — p-values are reported with R^2 values, but no test statistics or relevant model coefficients (effect sizes) are reported in the main text. I wonder if this is because the effect sizes are small? More transparency is needed to be able to evaluate these patterns. Further, the R^2 value estimated by phylolm is likely not interpretable in the same way that a standard coefficient of determination is, as the data are hierarchically structured (a violation of standard linear regression); see this paper (‘R^2s for Correlated Data: Phylogenetic Models, LMMs, and GLMMs’: https://doi.org/10.1093/sysbio/syy060). These comments reflect a general concern about the regression approaches employed here.

An alternative approach I might consider could be something from the field of machine learning, like random forests regression (with an adjustment for phylogenetic signal), which may provide a simpler assessment of how each of the 20+ predictor variables contribute to the overall predictive accuracy of a regression model (e.g. for predicting a given molecular rate). Given the scale of the data here, this might be very powerful. For that type of approach, as an alternative to model coefficients, one can easily extract individual variable importance scores, as well as an overall understanding of how much predictive information is contained within the predictors (e.g., AUC). These machine learning approaches also make essentially no assumptions about the form of the underlying data, and so may be well suited for the task here; no model selection or predictor selection is required (just give it everything and let the black box sort it out). This is an entirely different approach than what has been employed here, so I would leave it to the editor to decide if some exploration like this should be attempted. Nonetheless, I am so far not persuaded by most of the regression analyses that are reported, as they seem to imply small, though statistically significant, effect sizes.

#######
Below, I provide additional responses to responses that the authors provided in response to my initial review. For the five responses below (noted numerically), I have copied the author’s response and provided an additional comment.

1)
Author response:
Response. Residual variation in regression was indeed modelled using Brownian motion, with the amount of phylogenetic signal in the residual error modelled using the variance rate σ2 (this is equivalent to using λ, and the two parameters are non- identifiable; Ho and Ane 2014, doi:10.1093/sysbio/syu005). Values of σ2 were higher in regressions involving genome-wide molecular rates of dN, dS, and intergenic regions than in regressions of principal component loadings. This is intuitive, showing stronger phylogenetic signal in genome-wide rates than in the rates decomposed into subsets of lineages and genes. For interested readers, the data repository now points to the estimates of σ2.

Reviewer response:
There seems to be some confusion here. σ2 and λ are not equivalent parameters and are not used in the same context in phylogenetic comparative methods — the former is typically interpreted as the rate of character evolution, while λ is an estimate of a branch-length transformation parameter sometimes interpreted in the context of phylogenetic signal. In any case, values of σ2 are not appropriately interpreted as more or less ‘phylogenetic signal;’ these are different things. I suggest the authors may want to consult Luke Harmon’s textbook on these topics.

The paper that the authors refer to does not include the word ‘identifiable,’ nor would it make sense to try to estimate these parameters at the same time, so I do not follow the interpretation offered here. σ2 is not a measure of phylogenetic signal, it

is a parameter that estimates the rate of character diffusion or variance. In the context of regression, σ2 is used to account for variation in the residuals, assuming a pure BM process. If instead λ transformation is estimated/used, a pGLS will try to account for a mixture of BM and non-BM processes in the model residuals (this is described in the paper the authors reference above in the section 'Modified Brownian Motion Models.'

2)
Author response:
Response. Regression analyses have been added using the species-specific data rather than family averages, and these results are available in the supplementary material. We find that species-level regressions have far lower R2 values than family averages. This is expected given that molecular rate estimates correspond to the full history of molecular evolution along the lineages from each sampled individual back to the common ancestor shared with its sister family. In this sense, molecular estimates are not specific to the sampled species but more broadly represent an average rate across the history of the family. As such, the family-level trait average is far more likely than the species trait to be associated with our molecular rate estimates. This 'whole lineage' aspect of our molecular rate estimates is now explained in the text, as well as our comparison with species-level data.

Reviewer response:
I still do not fully agree with the authors' logic here. The "family average" is a problematic representation of the underlying trait data and does not truly correspond with the genetic samples in the dataset. As I suggested in my prior response, running the regressions with resampled trait values from within the family is perhaps more palatable but, even still, sub-optimal. An alternative approach emphasizing "family averages" could be to estimate an ancestral state for each family (a phylogenetic 'average'), prune the terminal lineages to the family MRCA age, and then use the estimated ancestral state values; this would attempt to avoid the sources of bias I am concerned about. In any case, a 'phylogenetic' average would be required to account for what the authors describe in their response, and not a standard mean.

Also, see my comment above about the interpretability of the R^2 value from phylolm; I suspect that a different approach is required if the goal of each individual regression analysis is to assess how much variance in the data is explained by the model (e.g., R^2pred, as noted in the paper by Ives).

While it is true that the molecular rate estimates as presented will reflect the average rate across the history of a branch (+/- some residual variation due to taxon sampling), it is also true that a biased taxon sample (in this case, a limited representation within families) is expected to lead to molecular (or any character) rate underestimates because the underlying substitution model rate matrix cannot 'see' the missing lineages (e.g., missing substitutions). It is not possible to know how much bias is introduced by using family-level trait averages without measuring or simulating it; I leave it to the editor to decide if further investigation is required, but I am not fully convinced by the authors' justification.

3)
Author response:
Response. Mass-corrected variables are used in the figure for visualization purposes, but this was not the case in regression analyses. Instead, body mass was included as a covariate in every regression analysis. This addresses its joint impact with other variables on molecular rates. For visualization purposes, variables were mass-corrected in cases with a strong association with body mass, such as in the case of generation length ($p < 0.001$, $R2 = 0.67$), beak depth ($p < 0.001$, $R2 = 0.65$), and brain size ($p < 0.001$, $R2 = 0.89$). In these cases, body mass has a strong impact on visualization of the ancestral states. In contrast, the correction has limited impact on the visualization of clutch size, which is not predicted strongly by body mass ($p = 0.02$, $R2 = 0.03$). The figure and caption have been revised to clarify these points.

Reviewer response:
The author's response does not address my original comment: What is a 'mass-corrected variable' in this context? The authors' response does not actually explain what they mean here. Does this mean that the data points in the figures are mass-corrected residuals from a regression model? Are they normalized by mass (e.g. divided by mass?)? Further, it is not clear what it means to use 'mass-corrected variables "for visualization purposes" — does this mean that the data points depicted in the figures are not the same as the data points used in the regression analyses?

4)
Author response:
Response. This is a particularly useful comment and has led us to implement a more sophisticated approach to multiple regression. We now consider every possible combination of variables in our multiple-regression models, and weight models by their AICc for model averaging. We report the model-average results across weighted models. The results are consistent with those of previous analyses with the single exception being the results for dN rates. In this case no variable consistently emerges as a strong predictor across all analyses. Traits that are found to be significant predictors in subsets of analyses are discussed further in the text, including beak depth, brain mass, and tarsus length. Some analyses that show such traits are those where all palaeognath taxa are included, or multiple-regression models based on single-variable analyses.

Reviewer response:
The model averaging procedure is not fully described in the methods; this is required to better understand what is happening here. For example, are the authors weighting the model coefficients by the alternative model weights? It's not clear.

5)
Author response:

Response. The recovery of deep branches in this species tree is likely to be affected by several forms of inference error. At this deep scale, model- misspecification error arising from true gene-tree discordance (i.e., incomplete lineage sorting) will be outweighed by gene-tree inference error. This is demonstrated by very low branch supports in individual genes at deep branches. Importantly, tree inference error is greater at deeper branches, while the error due to incomplete lineage sorting is similar across the tree even with fluctuations in population size. An excellent description of these relative amounts of error is given by Bryant and Hahn (2020, section 4.2, doi: hal-02535651).

Reviewer response:
This seems to be an argument in favor of excluding genes with low branch support on deep edges, as this would imply that the error in estimating branch lengths near those deep edges would also be high (or that there is very little signal of branch length information).

#####

Line-by-line comments:

Abstract: —
Line 18 ‘ we find that the dominant predictors’
— rephrase to say, ‘of the traits we evaluated’ …

Line 107 —
R2 is not a particularly useful metric without also knowing the model coefficients (particularly for multiple regression)—e.g., the effect sizes. A large R2 and a low effect size do not suggest a robust relationship. Also, see my prior comment about R2 for correlated data.

Line 110— same comment as above — what is the standardized effect size of the noted predictor in the optimal model?

Line 130—
Every time statistical relationships are reported (when a significant ‘predictor’ is being emphasized), I think it is important to report the estimated standardized effect size (e.g., the relevant model coefficients), the p-value, and optionally the appropriate coefficient of determination R2 (as suggested by Ives). In my opinion, relationships that have a low effect size (model coefficient) do not suggest a strong effect, even if the p-value is low.

Line 153-158
Can the authors clarify—lines 157-58—are the models testing body size set up as ‘y ~ body size + covariates’ — e.g., directly testing the effect of body size while including other covariates in the model? If so, what is the standardized effect size for body size, p-value, etc., as above?

Line 166
The suggested ‘lilliput effect’ is a hypothesis and not firmly established (e.g., through fossil evidence). Maybe should qualify it a bit more cautiously here.

Line 170-172
Other studies seem to be able to detect population size fluctuation with whole genome data (e.g., Stiller et al. 2024, and Houde et al. 2020). I suspect one reason it may not be detected here could be because the data are not analyzed in a coalescent framework. I have no strong opinion that this is a better way to analyze the data; however, now that it has been reported in two independent studies, the fact that it is not observed here suggests to me that the approaches used by the authors have introduced a source of bias/noise into the results (perhaps due to fixing the gene tree topologies to the species tree topology).

Line 182—
What is a ‘single gene rate averaged across lineages’ and how is it computed?

Line 185-187
Again, what are the effect sizes? — they are not reported in the main text. I suspect the relationship between GC content and rates may be due to model misspecification in genes with high GC content. This is now well documented in the literature — e.g., Stiller et al. 2024 exclude loci on this basis. Loci with higher GC content are more likely to violate model assumptions of stationarity, which could cause apparent rate variation (e.g. more rate variation across the entire phylogeny could look like a higher average rate). Non-homogeneous substitution models may be required to assess this more robustly.

Line 194-196—
This supports what I suggest above regarding model misspecification.

Figure 1—
Again, it is challenging to interpret the models without clearer reporting of the effect sizes. Line 202—‘mass corrected’ is again not clearly explained. Is this model residuals? Is this generation length/mass?

Line 220 —

How many components is < 5%? What is the effect size/test statistic? — just reporting the p-value does not seem sufficient.

Line 224—
The terminal branches popping out here are suspect to me and suggestive of some kind of artifact/bias. My guess is that it may be due to sequencing error, which I would expect to induce extensions of terminal branches.

Lines 236-245;
I am not sure we can rule out that these patterns are due to rate estimation errors at nodes with high levels of incomplete lineage sorting.

Line 285;
If this is a permutation test, what is the test statistic and its value/interpretation? It should be reported here.

Line 297-299:
Really cool result!

Line 303-305 — If true, this is the most compelling result of the paper, in my view.

Line 324— again, what is the test statistic etc.

Line 330, what is the effect size? I suspect it is extremely low, given the R2. In this case, the significant p-value is possibly a consequence of the massive sample size, suggesting very high sensitivity to extremely minor effects. I see that the approach with model averaging (not clearly explained) still includes tarsus length, which is encouraging; however, the effect size still seems to be extremely low. I suggest reframing in terms of model selection.

Line 334:
The authors refer to different tarsus 'optima' — but it does not appear that they have fit models of stabilizing selection (e.g., OU models) that would allow them to make this statement in a quantitative way.

Figure 4a
The results discussed here appear to show a very marginal effect. As I noted elsewhere, the effect size is not clearly reported.

Line 400:
I am not convinced that the paper has demonstrated this; it seems like the study shows a variety of significant correlations with low effect sizes.

Line 468:
These results look nice — the authors should report the effect sizes for each regression they report a p-value for. Including those values in supplemental data is not sufficient for transparency (though I have not been able to access these data tables). In any case, 'mass-corrected' is not clearly explained here, nor is it explained in the response to my prior comment.

Line 498:
Did the authors test for node density effects? It might be valuable to try to account for this directly, especially since the rate estimates are not coming from molecular clock models.

Line 506-509:
I remain concerned about forcing the tree topologies and the biases that this is expected to induce in rate estimates.

Line 520-526:
The concern here is not that the process may have occurred on a nearby branch — it is that the inference of a rate change might be entirely artifactual. I remain concerned about the choice to emphasize results derived from analyses of forced tree topologies.

Line 538:
"Continuous variables were verified to follow logarithmic scaling." — How did the authors do this? Typically, we log transform because it helps data conform to model assumptions, or because we believe that proportional scaling is more relevant than linear scaling. The quoted sentence means something entirely different.

Line 541-547:
I do not agree that a family average is always a good proxy of a phylogenetic average, which is what the authors are trying to estimate. These values may be very different; see below.

Line 548-550:
I do not agree that models with higher R2 values using family averages necessarily support this inference. Perhaps this would be true if this were based on effect sizes, but it is not clear if that is the case.

Line 551-557:
I am still not sure that this approach is statistically justifiable—the model averaging approach also does not directly address

my prior concern. The potential causal interactions among the examined traits mean that effects may only appear when multiple variables are included or excluded in particular models (a causal modeling framework like phylogenetic path analysis would be required to disentangle some of these interactions). Therefore, I am still wary of the approach that restricts multiple regression to only consider traits significant in 1:1 analysis.

Line 558-561:
Can the authors provide more detail about the logic of model averaging in this context? My prior comments about this did not request model averaging—they were about justifying the inclusion or exclusion of particular predictors on the basis of relative model fit.

Line 565:
Perhaps replace “Drivers” with “Correlations”

Line 573:
The authors again indicate model-averaging, but few details are provided as to how this procedure was carried out. I see that the description is included in line 990 about the MuMIn package but no detail is provided about what is being averaged.

Line 590-594:
It is still not clear to me that this procedure is statistically valid; the issue with mapping gene tree rates to species tree rates is not only about rates on a focal branch vs. a neighboring branch but rather about substitutions inferred on a focal branch that does not actually exist.

Referee #2

(Remarks to the Author)
I think the manuscript drastically improved and now incoporates GC content as one of the most important covariates of molecular rates. I happy how the authors addressed my concerns, and I think in the current form the paper will set a bar on comparative genomic approaches.

(Remarks on code availability)
I reviewed some of the code

I noticed that one of the scripts refers to a local library/file source("~/Dropbox/Research/code/brsPerTslice.R")

I think these should be avoided. The code should be able to run on its own.

Referee #3

(Remarks to the Author)
This is a revised version of a manuscript I previously commented on. The authors have made substantial changes in the new version and have sufficiently addressed previous comments.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
In this revised submission, the authors have graciously and thoughtfully responded to all of my prior comments, and I thank them for their efforts to improve the manuscript. The addition of Bayesian regression and random forests analyses is impressive and improves the statistical rigor of the manuscript. These techniques replace previously opaque model averaging, so this is in the right direction. Slightly more detailed reporting of the random forests analyses would be welcome (particularly with respect to model performance). The authors note that they assessed model performance with RMSE, but those details do not appear to be reported. Were there any other assessments of the random forests performance?

While I do not agree with all of the provided explanations to my prior questions (principally those related to the emphasis of a fixed tree topology), this perhaps stems from differences of opinion, and I am happy to co-exist with differences of opinion. I am generally satisfied with the other responses offered by the authors, and the paper has been significantly improved through review. Given the brevity of their comments, the other two reviewers do not appear to have reviewed the revised manuscript carefully; considering the substantial changes in the latest revision, it may be important to request additional reviews from new reviewers. However, should the minor changes I suggest here be incorporated, I don't think I need to see the manuscript again (at the editor's discretion).

Here are a few minor other comments for the next and hopefully final revision:

Since I read the last revision, there is a new pre-print linking rates of genomic evolution to tarsus length in birds https://www.biorxiv.org/content/10.1101/2024.04.30.591925v1 -- the authors may want to cite this pre-print where appropriate (though extended discussion of a pre-print is probably not warranted).

The paper could also be improved with an enhanced/expanded discussion that more thoroughly compares their new results to the results of prior studies (some recent), specifically in the conclusion section. This is important and should be the last issue that the authors need to address (in my opinion).

For example, this sentence appears in the conclusion:
"These analyses show that the key drivers of genome evolution in birds include clutch size, generation length, and tarsus length – traits long considered important in avian life history and ecology – along with GC content in genes."
All of these factors have been previously investigated with respect to patterns of avian molecular evolution, and relevant works should be cited here (some are cited elsewhere in the paper). Any appropriate commentary (as judged by the authors) can also be added to help readers place these results within the broader research context of these topics. For example, the recent work by Berv et al (2024) https://www.science.org/doi/10.1126/sciadv.adp0114 suggests a weak link between clutch size and the mode of molecular evolution (as distinct from (though related to) the rate) -- and it is interesting that clutch size also comes out as an important factor in this analysis of a much larger genomic dataset (among other interesting/contrasting patterns). As the authors are aware, many other papers have investigated various life history traits and patterns of molecular evolution, and the conclusion section could be better used to emphasize how these new analyses build on prior results and the work of other researchers (including, for example, these important works: https://doi.org/10.1093/molbev/msx236, https://www.science.org/doi/10.1126/science.aat7244, https://www.cell.com/current-biology/fulltext/S0960-9822(17)31081-3, which investigate similar questions using model-based approaches). It may be especially important to emphasize how these new results differ from any prior scientific consensus (for example, the proposed associations with body mass). I suggest that the authors carefully review their concluding statements in the context of prior work and add more detailed acknowledgments. I know that Nature has strict length limitations, but I would like to ask the editors to give the authors some grace in including additional contextual details that will improve the scholarship of the work.

UNIVERSITY OF COPENHAGEN
DEPARTMENT OF PUBLIC HEALTH

Senior Editor, *Nature*,

1 MAY 2024

Please find below our responses to reviewers for our manuscript “Drivers of avian genomic change revealed by evolutionary rate decomposition”. This provides detailed explanations for the extended changes made in this version, as recommended.

Yours sincerely,

David A. Duchêne and Simon Y.W. Ho, on behalf of all co-authors

David A. Duchêne
Associate Professor
Department of Public Health
Øster Farimagsgade 5
1353 København K
+45 35 32 10 84
david.duchene@sund.ku.dk

Editor's comments

I want to apologize again for the delay in our decision on your manuscript, "Drivers of avian genomic change revealed by evolutionary rate decomposition", which has now been seen by 3 referees, whose comments are attached below. While they find your work of potential interest, as do we, they have raised important concerns that in our view need to be addressed before we can consider publication in Nature.

Response. We thank the reviewers and editor for their valuable comments, which we have carefully addressed in this letter and revised manuscript.

Beside the technical comments of the referees, which of course must be addressed in full, we also want to highlight the need to improve the organisation and accessibility of the manuscript. Please keep in mind the broad readership of Nature. As currently written, we do not feel that the paper is accessible to readers outside the field of evolutionary genetics.

Response. The text has been revised extensively to improve accessibility to a wider audience. Specifically, much of the terminology has been simplified, such as that relating to molecular clocks and evolutionary rates.

Please also update us on the status of the preprint describing the ClockStaRX method. It is a concern that a large proportion of results described here rely on an unpublished method. Further, please do make sure that - if the preprint is approaching publication - there is no overlap in results presented there and in your submission with us (concern of potential dual publication).

Response. The article on ClockstaRX has now been published in *Genome Biology and Evolution* (doi: 10.1093/gbe/evae064). None of the data or results overlap between the two manuscripts. Examples on avian data only involved older data (from 2014), and have since been removed. The ClockstaRX paper validates and demonstrates the capabilities of the software using simulations.

Referees' comments:

Referee #1 (Remarks to the Author):

Review for: Drivers of avian genomic change revealed by evolutionary rate decomposition

The presented work by Duchêne et al is a fascinating exploration of recently available whole genome data for birds. In general, I really enjoyed reading this

paper, and I think it will interest a broad audience. It is exciting that we are finally moving into an era in which this level of detailed analysis can be attempted at such a wide taxonomic scale, further extending our reach into the past. There are many detailed and nuanced analyses reported here, so it will take some time and dedication for any reader to fully appreciate them. It is possible that some of my comments are a consequence of details I missed; I apologize to the authors in advance if so.

Response. We thank the referee for the positive assessment. This version of our manuscript has been substantially reworked to improve accessibility. We hope that the primary insights from our work are now easy to grasp for the broadest scientific audience.

I have read through the paper several times, and I think I have been able to follow the primary narrative. The paper has two parallel analytical tracks; the first is to examine how a suite of traits is related to different aspects of molecular rate variation. I was very impressed by the evaluation of 29 traits, which is indeed comprehensive. The second track is to "decompose" different aspects of molecular rate variation across individual genes and phylogeny branches to assess which, if any, have "an unusually large impact on rate variation." The authors employ a technique implemented in the software ClockStarRX, developed by the authors. This technique is available, though it currently exists as an unreviewed bioRxiv preprint.

Response. ClockstaRX is has now been accepted for publication in *Genome Biology and Evolution* (doi: 10.1093/gbe/evae064), and is cited as such in this manuscript.

I have some thoughts, comments, and concerns on both tracks, so I will split my review into two sections. Most of my comments are related to the methods associating traits with molecular rates. Then, I will provide some more minor comments. While this paper and overall approach have tremendous merit, addressing the issues I highlight below will significantly strengthen the paper's conclusions.

Linking trait variation to individual patterns of molecular rate variation:

The authors' approach is to first examine each of a set of 29 traits individually with respect to each aspect of molecular evolution (dN, dS and w), using "simple" phylogenetic regression.

1) I assume this means that the regression model used assumes residual variation follows a Brownian motion process (I believe this is the default for phylolm)? Not

clear what model is used for these regressions. If so, I suggest the authors specify the methods, as it is a key assumption.

Response. Residual variation in regression was indeed modelled using Brownian motion, with the amount of phylogenetic signal in the residual error modelled using the variance rate σ^2 (this is equivalent to using λ, and the two parameters are non-identifiable; Ho and Ane 2014, doi:10.1093/sysbio/syu005). Values of σ^2 were higher in regressions involving genome-wide molecular rates of d_N, d_S, and intergenic regions than in regressions of principal component loadings. This is intuitive, showing stronger phylogenetic signal in genome-wide rates than in the rates decomposed into subsets of lineages and genes. For interested readers, the data repository now points to the estimates of σ^2.

2) Figure 2 indicates several variables that are "mass corrected," but no details about this are provided in the methods. Why is mass-correction important in this context? Why was clutch size not mass-corrected if the other highlighted variables were? This is an important point to clarify because clutch size is highlighted as the most important predictor of molecular rate variation. Revell 2009 ('Size-correction and principal components for interspecific comparative studies') provides useful guidance.

Response. Mass-corrected variables are used in the figure for visualization purposes, but this was not the case in regression analyses. Instead, body mass was included as a covariate in every regression analysis. This addresses its joint impact with other variables on molecular rates.

For visualization purposes, variables were mass-corrected in cases with a strong association with body mass, such as in the case of generation length ($p < 0.001$, $R^2 = 0.67$), beak depth ($p < 0.001$, $R^2 = 0.65$), and brain size ($p < 0.001$, $R^2 = 0.89$). In these cases, body mass has a strong impact on visualization of the ancestral states. In contrast, the correction has limited impact on the visualization of clutch size, which is not predicted strongly by body mass ($p = 0.02$, $R^2 = 0.03$). The figure and caption have been revised to clarify these points.

2a) A lesser point is that clutch size is probably not optimally modeled as a continuous trait (it is a type of count data). Although many authors do this, it is better modeled with a Poisson or binomial regression for count data. These can be implemented using the BRMS R package while accounting for phylogenetic signals. In most situations, approximating clutch size as a log-transformed continuous trait is probably fine, but I emphasize this here because clutch size is noted as the most significant feature in the set of traits; thus, this part of the paper hinges on the validity of this result.

Response. Our response variable in regression is consistently molecular rate. Poisson or binomial regression is relevant when the response is count or binomial data, which is not the case in our study. Clutch size and categorical traits are treated as independent (explanatory) in every model. This is now clarified in the Methods section.

3) I am concerned about the statistical validity of taking the variables that are identified as being significant in 1:1 regression analysis (e.g., clutch size vs w) and then using them together in multiple regression. The intention of this latter aggregation in multiple regression is not clearly articulated; the authors say, lines 487-490: "We then ran multiple phylogenetic regression, where all the significant explanatory traits from the individual regressions were added as explanatory variables as partial terms, testing whether significant variables explained molecular rates above the effect of other traits." I think this is saying that the tests are designed to evaluate whether a focal dependent variable (e.g., clutch size) remains significantly associated with the independent variable (e.g., w) after controlling for the effects of other variables identified in 1:1 regression. If this is the correct interpretation, it could be more clearly stated. This may be a fine thing to do, but as no examples of model comparison are provided (e.g., comparing simpler to more complex models through their likelihood or AIC), it is challenging to assess the justification for running the multiple regression models. In general, it seems preferable to include multiple predictors when their inclusion is justified in terms of relative model fit to the data or when there are stated hypotheses about multiple or interaction effects.

Response. This is a particularly useful comment and has led us to implement a more sophisticated approach to multiple regression. We now consider every possible combination of variables in our multiple-regression models, and weight models by their AICc for model averaging. We report the model-average results across weighted models. The results are consistent with those of previous analyses with the single exception being the results for d_N rates. In this case no variable consistently emerges as a strong predictor across all analyses. Traits that are found to be significant predictors in subsets of analyses are discussed further in the text, including beak depth, brain mass, and tarsus length. Some analyses that show such traits are those where all palaeognath taxa are included, or multiple-regression models based on single-variable analyses.

Moreover, it is unlikely that these traits evolve independently; they will be correlated and will follow a complex multivariate variance-covariance structure. I suspect the authors know this, as noted above, there are some variables that are qualified as being "mass corrected" (though this is not explained). Multivariate phylogenetic regression techniques are available in R packages like mvMORPH (see mvgls) or SLOUCH, for example, and could be employed here to assess how

the multivariate relationships among these traits are related to various evolutionary rates. I recognize this is a somewhat distinct research area and do not ask the authors to explore it (unless they want to); only to better acknowledge that focusing on individual traits may be misleading when traits covary in complex ways.

Response. Our response variable in all cases is univariate molecular rate, while covariance among independent variables is taken into account in multiple regression. The molecular rate variables follow distinct evolutionary processes, so we propose separate hypotheses and models to explain each, instead of handling a multivariate response (e.g., d_N is heavily affected by selection-related variables, while d_S is affected by those affecting germline-replication).

We wish to emphasize that multiple regression accounts for the joint impact of explanatory variables on the response. This is why the parameter estimates for covariates are also known as the effect after other variables have been 'partialled out' or 'controlled for'. The inherent structure of the linear model accounts for any correlation among variables. One important example is body mass, which was included as a covariate in all multiple-regression models.

4) I was surprised by the result that body mass is apparently not related to any metrics of molecular evolution, as this has been reported in numerous other studies, including those which employ genome-scale data (e.g., Botero-Castro et al 2019; 'Avian Genomes Revisited: Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds'). [e.g., Lines 149-151, I assume this means all types of molecular rates and not only w? Apologies if I have misunderstood]. It could be useful for the authors to speculate further on what might explain this result. This is especially surprising in the context of 1:1 regression analyses that evaluated body mass and did not control for covariance with other traits (which I would typically guess would remove the effect of body mass). This seems somewhat in conflict with results reported in the associated paper (Stiller et al.), which report an increase in substitution rates associated with the end-Cretaceous transition and an associated decrease in body mass. The paper reports (Fig 2b) that "log(mass-corrected generation length)" is associated with substitution rates, but not if "log(generation length)" is associated (hence the importance of clarifying the issues of mass correction noted above). log(body mass) correlates well with log(generation length) in birds, and since generation length is detected as a significant predictor, this pattern is indeed perplexing. It is also somewhat paradoxical, as body mass is expected to be a reasonable proxy of effective population size through the relationship body size has with environmental carrying capacity. It would be useful for the authors to check some simple metrics of this expected association, for example, checking if log(root to tip branch length) [from the gene trees] is related to log(body mass) (with both phylogenetic and non-phylogenetic regression).

Perhaps another option is to be more circumspect about this result in light of other papers that use more sophisticated modeling approaches to detect associations (below).

Response. Body mass was significantly associated with genome-wide molecular rates, as expected, and was therefore present as a covariate in multiple-regression models. In multiple regression, body mass did not significantly explain molecular rates, pointing to a greater impact of other variables. We apologize for this lack of clarity, and have revised the manuscript accordingly.

5) The finding for beak depth is fascinating; it would be helpful for the authors to briefly clarify what this feature is (in context), as it may not be apparent to non-bird experts. Lines 128-131: I am surprised by the inclusion of hummingbirds on this list, as I generally think of them as having relatively long, high aspect ratio beaks relative to their body size, but maybe this is just me misunderstanding what 'beak depth' means. I suspect this makes more sense if we are talking about beak depth relative to body size, which seems to be the focus (in Fig 2). Could beak depth be a proxy for body size (if not mass)?

Response. Beak depth is the distance or height from the the start of the beak at the top of the head, to the start at the bottom of the head. Deep beaks are often associated with adaptations for breaking hard objects, as seen in some parrots and finches. Very shallow beaks are often seen in birds that reach into small areas or that are nectarivores, as observed in many hummingbirds. Beak depth was not found as significant in the new model-averaged regression explaining d_N, and instead we mention the several variables that in appear as significant in multiple-regression models informed by single-variable analyses (including beak depth, brain size, and tarsus length). We have clarified these points in the manuscript and have modified Fig. 1 accordingly.

6) In general, although the huge scale of the work justifies simplifying analytical assumptions, the authors should probably acknowledge as a caveat somewhere that the 'simplified' regression approach makes many assumptions that can lead to biased inference. Although it would be impossible to use on a dataset of this scale, alternative multivariate approaches should be much less biased and more sensitive. One example is the model implemented in the software 'coevol' (Lartillot and Pujol 2011); this could be used to spot-check some results.

Response. As stated by the reviewer, the scale of our genomic data is too large for Bayesian analysis (e.g., as implemented in coevol) to be feasible, and the focus on simple regression in coevol does not allow our primary aim of disentangling the joint impact of multiple variables. In other words, the insights drawn would be comparable to our simple regression analyses. The addition of exhaustive multiple-

regression model-fitting and model averaging addresses the simple approach taken in the previous manuscript, adding robustness to our results.

Linking trait and lineage variation to decomposed aspects of principal components:

1) I struggled a bit more with these sections, as they make use of an analytical pipeline employed in the ClockStaRX software, currently reported in a bioRxiv preprint (https://www.biorxiv.org/content/10.1101/2023.02.02.526226v1.full). The inclusion of Extended Data Figure 1 was very helpful to understand exactly what is happening here. It appears that there is some preliminary overlap between analyses presented in the preprint and those reported in the current manuscript, as the preprint investigates some similar patterns in avian phylogenomic data.

Response. The analyses using avian data were excluded entirely from the ClockstaRX paper. Those analyses had been based on an older data set (from 2014) with far fewer taxa and genomic regions. The ClockstaRX paper includes only simulations (doi: 10.1093/gbe/evae064).

In the main text section introducing these results [Line 184], I think it would be helpful for the authors to insert one additional short sentence providing more context for what this analysis is doing. The current language was not sufficient for me on my initial read, and I had to go to Extended Data Fig 1 to understand that the input data matrix into the PCA is a matrix of branches (columns) and genes (rows). I inferred that these analyses were performed independently for dN, dS, and w? It is not clearly stated if these parameters were analyzed jointly or independently. Further, can the authors clarify if the underlying PCA in ClockStaRX is a Phylogenetic PCA? I am not sure if it is necessary in this context, but as these data are all inherently comparative, it may be appropriate.

Response. The PCA employed is not phylogenetic. The variables used correspond to all tree branches, such that the new dimensionality accounts for the full evolutionary history. This is now clarified in the main text and methods. As an analogy, our approach can be compared to using phylogenetically independent contrasts, where non-independence is addressed by using the ancestral contrasts among each pair of sister lineages. Our approach nonetheless makes the overarching assumption of the accuracy of the species tree and its concordance with gene trees, which we now make explicit in the Methods.

2) In general, I found these results to be interesting, and compelling and they convinced me that this approach might be a good way forward for tackling some of the gene-phenotype associations observed at this broad scale. While most conclusions seem sound and well justified, I am concerned about one important

methodological choice and its impact on some of the major results. The authors note in the methods that all gene trees were estimated on a fixed topolgoy indicative of the species tree (thus, they only vary in branch lengths). This was done so that more data could be evaluated for deeper nodes, as fewer and fewer concordant edges will exist as we go back in time. While I appreciate this fact, the downside of this choice is that it may introduce bias, resulting in some of the exact patterns that the paper reports (e.g., fast rates of molecular evolution associated with the origins of ancient clades, exactly where we expect there to be a lot of gene-tree discordance). The authors cite Mendes and Hahn (39), so they are aware of this potential source of bias. I personally believe that these patterns are likely real, but because the authors do not directly account for gene-tree topology discordance here, these results could be an artifact of incorrectly mapping substitutions to edges that do not exist in the true gene trees. As a solution, I wonder if it would be particularly challenging for the authors to provide as supplementary data a small vignette corroborating one or two of the interesting edges that are detected (e.g., the stem branches of Telluraves or the new Elementaves) with analyses that allow the gene tree topologies to vary? This would help give confidence that these results are not artifacts or biased in the vein of Mendes and Hahn.

Response. The recovery of deep branches in this species tree is likely to be affected by several forms of inference error. At this deep scale, model-misspecification error arising from true gene-tree discordance (i.e., incomplete lineage sorting) will be outweighed by gene-tree inference error. This is demonstrated by very low branch supports in individual genes at deep branches. Importantly, tree inference error is greater at deeper branches, while the error due to incomplete lineage sorting is similar across the tree even with fluctuations in population size. An excellent description of these relative amounts of error is given by Bryant and Hahn (2020, section 4.2, doi: hal-02535651).

In addition, the effect of substitutions artefactually appearing in assumed branches (‘substitutions produced by incomplete lineage sorting’, or ‘SPILS’ as termed by Mendes and Hahn, 2016, doi: 10.1093/sysbio/syw018) is local. This means that the actual acceleration/deceleration of rates will have happened at branches that are adjacent to those branches that are assumed (and those with similar timescales). We have extended our discussion to acknowledge these points more explicitly.

3) I loved the discussion of the chromosome-scale variation, the importance of microchromosomes, their gene content, and the recombination landscape. Great stuff: this is really important!

Response. We thank the reviewer for the positive assessment. In addition to the analysis of chromosomes we have performed analysis of GC content, which is

another intuitive and important predictor of hotspots of gene expression and recombination.

Minor comments/questions by line:

Line 53:
The authors may wish to acknowledge body mass (Weber et al. 2014, Botero-Castro et al. 2019 (body mass), Berv and Field 2018 (body mass, metabolic rate) (cited elsewhere in the manuscript)), even though their current results do not support it.

Response. These references have been incorporated into our text, as suggested.

Line 64-65:
Accelerated rates could be indicative of relaxed purifying selection in coding regions; as written, this phrase sort of implies the opposite.

Response. We now refer to the possible rate acceleration due to relaxed selective constraints.

Line 71:
"Modelling" is spelled with two l's — there are a few other spots where this is done. Not sure if it is a typo or British spelling.

Response. We have checked our manuscript carefully to ensure that British (Oxford) spelling is used throughout.

Line 102:
Generation length is correlated with which type of molecular rate? Not clear.

Response. This sentence has been rephrased to make clear that it refers to the link between generation length and rates at ingergenic regions.

Line 134-35:
Were any other taxonomic exclusions made based on extreme values?

Response. The two palaeognath taxa with extreme values of wing shape were the only exclusions (Struthionidae and Rheidae), and this is now made clear in the text.

Line 158:
Can the authors clarify, w is not associated with any traits in the dataset?

Response. The manuscript now makes clear that genome-wide ω was not associated with any of the traits in the data set, either in simple regression or model averaging analyses.

Line 164-71:
The proposed explanation with respect to population size does not seem consistent with results reported in the other submitted paper (Stiller et al.), which report lineages associated with population size expansion associated with the end-Cretaceous extinction. Surely, the impact of Ne varies across the genome, but it's not clear how this might impact the overall proposed associations with body size.

Response. Since we explore the rate of the ratio of nonsynonymous to synonymous substitutions, our inferences are entirely different from those of Stiller *et al.*, which were based on the ratio of coalescent units (measured from sites across the genome) to time. Importantly, our analyses treat ω as an indirect measure of population size, also affected by selection coefficients and mutation rates.

Specifically, what our results suggest is that the impact of population size on genome-wide selection was either relatively minor, or can no longer be detected at the deep timescales explored. Gene-by-lineage interactions are likely to play an important role in obscuring these effects. Other factors that contribute to obscure the genome-wide effect of population size include phylogenetic error, substitution saturation, and incomplete lineage sorting. This paragraph has been revised to include these points.

Line 210-214:
How can this be true if there is no association between body size and variation in molecular rates? Seems like a paradoxical argument.

Response. We did find a strong association between body mass and molecular rates in simple regression as expected, although this result was no longer significant in multiple regression analyses. In addition, our result is not inconsistent with major variations in population size surrounding the K-Pg transition, given that we did not perform direct estimates of ancestral population size.

Line 215:
The authors may want to cite Brown et al. 2008 (Strong mitochondrial DNA support for a Cretaceous origin of modern avian lineages) and Berv and Field 2018 (Genomic Signature of an Avian Lilliput Effect across the K-Pg Extinction), who also investigated this idea. Berv and Field 2018 is cited elsewhere, so it would not increase the reference count to acknowledge that idea here.

Response. These relevant works are now acknowledged, as suggested.

Fig 2 caption; line 220:
It would be helpful to be specific about "locus rates" — e.g., that this means dN, dS, and w (for readers who focus on the figures).

Response. The figure and caption have been revised to make it clear that the data correspond to ω rates.

Line 272:
The phrase "under high mutational pressure synteny" - I think there is a missing comma here: "under high mutational pressure, synteny"?

Response. This mistake has been corrected.

Line 290:
The reported relationship between w and tarsus length is pretty tenuous; the effect size is low (R2=0.064). With this type of regression modeling framework and the huge sample sizes used here, it is not very difficult to get a "significant" result. It may be helpful to know if including tarsus in a model predicting w significantly improves model fit (through something like likelihood or AIC).

Response. As part of the newly added model averaging framework, we now show using AICc that the model including tarsus length far outcompetes the alternative null model (sum of weights of 1 for models including tarsus length). The methods have been updated accordingly.

Line 306-307:
As noted above, I am concerned that these observations of multiple early bursts are artifacts of the choice to fix the gene tree topologies to the species tree topology (even though I think they very well could be real).

Response. The text is now updated to clarify that our inferences refer to the given branch as well as neighbouring branches in cases of gene-tree discordance. In that respect, our result is an average across the data. The text is now more cautious regarding statements about this particular topology, since we are aware that multiple trees receive similar levels of support (shown in Stiller *et al.* 2024).

Line 308:
'Accelerated rates' - is this referring to w?

Response. The sentence is rephrased to make our reference to ω clearer.

Line 310-313: very cool result — I believe patterns of development/growth could be implicated along these deep branches and these results support that idea.

Response. We thank the referee for the positive assessment. We have refrained from mentioning the impact on development directly, but agree that our results point towards changes in gene regulation in early stages of life.

Line 320:
"signalling" see above; possible typo

Response. We have checked our manuscript carefully for consistent usage of British (Oxford) English.

Line 476-483:
Please be explicit about any data transformations, scaling, etc.

Response. A statement describing our approach to data transformation has been added to this paragraph.

Line 484:
(a less minor point)
Using the family-level averages for phylogenetic comparative data/analyses is probably not a good practice, as this can bias estimates of the evolutionary parameters estimated in the phylogenetic regression models. E.g., if the family level average is not close to the family level MRCA (which it very well may not be, depending on clade age and phylogenetic structure within families), the model may have to explain trait variation for a given terminal, depending on the direction of the bias. In general, in this situation, it would be optimal to use the actual tip-level species data and then resample different species data from within each family to see if the results are robust (though, I would advise the authors use actual tip-level species data and not to bother with resampling, as this would already be an accurate representation of the terminal trait data with respect to the genomes underlying those terminals). This is particularly important here as there is some attempt to associate these traits with underlying patterns of genetic variation. As is, it is difficult to know what kinds of bias this may have introduced.

Response. Regression analyses have been added using the species-specific data rather than family averages, and these results are available in the supplementary material. We find that species-level regressions have far lower R^2 values than family averages. This is expected given that molecular rate estimates correspond to the full history of molecular evolution along the lineages from each sampled individual back to the common ancestor shared with its sister family. In this sense, molecular estimates are not specific to the sampled species but more broadly

represent an average rate across the history of the family. As such, the family-level trait average is far more likely than the species trait to be associated with our molecular rate estimates. This ‘whole lineage’ aspect of our molecular rate estimates is now explained in the text, as well as our comparison with species-level data.

Line 511-512:
Related to my prior comment. Although the terminals here are sampled to represent avian families, it may be preferable to describe this as loadings of species. But, given the way the trait data are averaged, I am not sure.

Response. Following from the explanation above, our rate estimates are best interpreted as being averaged throughout the evolutionary history of each family, rather than being representative of individual species. Parent-offspring trios offer direct estimates of species-specific instantaneous mutation rates (e.g., doi:10.1038/s41586-023-05752-y), while our analyses offer a longer-term perspective of substitution rates.

Referee #2 (Remarks to the Author):

A) The manuscript "Drivers of avian genomic change revealed by evolutionary rate decomposition" attempts to identify life-history traits that correlate with rates of genomic changes in avian species. Specifically they look at protein evolutionary rates as well as intergenic regions. For their large scale analysis the authors rely on two datasets, one from Feng et al (2020) for the alignment data and Stiller et al (unpublished) for the tree data. The latter paper has been attached as supporting information and as far as I understand has been submitted else well. The authors complement these two datasets with trait data which they have collected from databases. The information for this are included in a spreadsheet available in a repository, and many of them rely on Tobias et al, 2022. Please note that the Feng data uses 1Kb pieces of intergenic sequences rather than whole genome alignments. While the underlying data has been used from other sources, the main innovation in this work is a decomposition method that would incorporate gene trees and species trees. They then use permutation tests to assess statistical significance.

B) Generally I think this is an interesting piece of work. It uses a large amount of data and attempts to identify general drivers of molecular evolution in avian species.

What I particularly like:
- Extended Figure 1 explaining their approach

- The speculation on tarsus length as an indicator of how panmictic/moveable individuals of a species are

What I think is missing:
- more focus on non-coding elements
- including variation into the genetic analysis (such as heterozygosity)

Response. We thank the referee for the encouragement and for providing a very helpful assessment. We agree on the value of adding data beyond the coding regions and their function. Accordingly, this version includes analyses of GC content and location along chromosomes.

C/D/F) What I think is important to consider and which I could not find info on

1) How do you resolve when a gene tree and species tree disagree?

Response. The method implemented in ClockstaRX identifies the gene-tree branches that are also present in the species tree (i.e., concordant branches). However, in an ancient system as that of avian taxa, this approach led to vast amounts of gene-tree branches being excluded from analyses. Since many genes will not have reliable signals of deep relationships (of either the species tree or gene tree), we opted to assume that all gene trees are identical to the species tree. The manuscript is now more explicit about the consequences of this assumption, and particularly when referring to events at specific branches of the avian tree.

2) GC content as an important variable is considered in the text, however it is not really tested for. Principally this could easily be incorporated into the study. So for example how do the top 20% of genes (e.g. Fig4) compare regarding their GC content relative to the rest of the genes.

Response. This is a very valuable suggestion that we have taken onboard. We have calculated GC content across loci and explored whether locus rates and principal components are associated with this metric. The manuscript and figures have been modified accordingly.

3) I think DNA methylation driven mutation rates get little attention. If GC content is high, so will be the CpG content which leads to increased mutation rates.

Response. The newly added analyses point towards GC content as an important mediator of mutation rates. GC content interacts with gene length and with distance from chromosome endings to explain gene molecular rates (averaged across branches). At high GC contents, short genes and those near chromosome endings have the highest molecular rates. Intriguingly, this is only the case for coding

regions, and particularly synonymous rates, pointing to mutational pressure at highly expressed regions (and possibly methylation, as suggested). This is an important addition to our understanding of genome-wide drivers of molecular rates, and is now presented prominently in the manuscript and added to Fig. 1.

4) It has often been noted that GC content might not be at equilibrium in birds. This has drastic consequences if not incorporated into the models of evolution which usually assume a stationary base composition. In particular for long branches and synonymous sites this could be an issue. Codeml does have models to test for this.

Response. The approach used for inference of molecular rates focuses on distance estimates for each taxon pair. This addresses the issue of non-stationarity to a large extent, since model parameters are inferred individually for each taxon pair. This was implemented in codeml, as suggested. We do not perform formal tests of equilibrium since these can lead to rejection of the data even under conditions in which unequal GC contents are unlikely to mislead evolutionary inferences (Duchene et al. 2017, doi:10.1093/molbev/msx092). We have revised the Methods to describe the flexibility of our approach in more detail.

5) For fig 3 - what is the proportion of genes that are in the same chromosomal location versus those that are not (e.g. due to intrachromosomal rearrangements)

Response. This raises the option of inferring rates of gene rearrangement and chromosomal translocation. Avian genomes have highly conserved synteny (doi:10.1126/science.1251385), such that this question is only moderately interesting in the case of birds. The question is also very different from that of nucleotide evolution. There are several metrics for non-sequence evolutionary rates, which are not all well understood and are difficult to model. For instance, translocations among different-sized chromosomes could be considered ‘faster’ evolutionary change than translocations among similar-size chromosomes, but how to interpret these changes is not clear. We agree this is an interesting avenue of future research, and we have added these topics to our conclusions.

6) How did you deal with chromosome number

Response. Data from chromosome rearrangement cannot be incorporated into existing well-understood evolutionary models, nor into the ClockstaRX framework of analysis. It is a form of complexity that might have some impact on our results, but is not currently tractable to incorporate explicitly into our analyses. Our study focuses on evolutionary rate at the nucleotide level, which is grounded in large bodies of theory, including those of DNA and codon substitution, nearly neutral evolution and molecular clocks.

Aspects of genome evolution such as genome size, karyotype, chromosomal translocation, gene rearrangement and transposition, and DNA 3-dimensional structure are promising avenues for future work. We now mention some of these topics in the conclusions as interesting future avenues.

7) Is there any analysis that considered ends of chromosomes, usually ends of macrochromsomes are very similar to microchromosomes

Response. We thank the reviewer for raising this topic. We have added tests of rates across chromosomal location. Specifically, we test whether distance to the ends of chromosomes is associated with molecular rates or with principal components (rates in subsets of lineages). Our data show a strong association between distance from chromosome endings and synonymous molecular rates in genes, mediated by GC content. This highlights the location of genes and their composition as dominant predictors of molecular rates, and these points are now presented prominently in the main text.

8) For fig 4 you identify many RNA related genes (ribosomes, splices

General comments:

- I think it would be good if the number of sites analyzed could be considered. For example in Fig 3 you mention number of coding and n intergenic. Is this the number of codons/genes/blocks - how many actual sites where included?

Response. We have added locus length as a variable to assess for evolutionary rates, and reveal an strong interaction between gene length and GC content in explaining mutation rates of genes. This has now been incorporated into Fig. 1. The data in Fig. 3 refer to the number of loci, and this has now been made clearer in the figure and caption.

- What has been done if there is variation in the chromosome number, such as extra chromosomes - in general karyotypes are conserved, but some variation is still present (e.g. falcons)

Response. As explained above, our focus on nucleotides follows extensive modelling theory on the topic and addresses one of the most fundamental forms of evolutionary rate. However, other measures of genome evolution are also interesting, and are now mentioned in the conclusions.

- How did you proceed with chromosome 1 (I assume you used a pseudo 1 for songbirds)

Response. As explained above, there are many measures of non-sequence evolutionary rates, so we have decided to focus our study on the fundamental level of nucleotides. Analyses at the nucleotide level do not consider chromosomal rearrangement, resting only on the assumption that nucleotides are homologous. We now mention other related fields of genome evolution in our conclusions.

- *There is no mentioning of the GRC and potential duplication outsourcing of genes in species that possess it?*

Response. Our study does not explore the evolutionary rates of chromosome size or presence/absence. The germline restricted chromosomes of songbirds are a good example of chromosome evolution by function, as are micro/macrochromomes and sex chromosomes. These are substantial departures from molecular rate theory at the nucleotide level, so they are not considered in our study.

- *where non coding RNAs considered at any point?*

Response. RNA sequences were not considered in this data set. However, this is yet another aspect of genome evolution that could be explored in future research.

- *Did you have any look at TE elements as genomic predictors (as these stopped spreading at some point)*

Response. Loci were not tested for the presence of transposable elements. The spread of these are also an interesting area, but a major departure from our exploration of genome-wide patterns of evolutionary rates at the nucleotide level.

- *Have you tried thinning the data? Your point estimates for each branch will get worse if you add more species. I understand you used a lower threshold for excluding genomic elements - but I think fewer species will perhaps allow you more meaningful fun omega ratios*

Response. Given the very long timescales considered, it is likely that our taxon sampling is sufficient for estimates of molecular rates to be meaningful. In fact, increased sampling will lead to identifying more ‘hidden’ substitutions (doi:10.1111/j.1558-5646.2007.00188.x). The dominant factor driving this signal are gene overall rates, which vary greatly and overshadow any variation in rates arising from taxon sampling (Duchene et al. 2020, doi:10.1093/molbev/msz291).

Line 375 form = from

Response. This mistake has been corrected.

Referee #3 (Remarks to the Author):

Summary of the key results:

This study utilises available genomic data, and morphological, ecological and life-history data, from 281 species of birds representing 87% of avian families. The genetic data come from the B10K community (Fang et al.) and the species attributes from the literature. The genomes are aligned and sequence variation between lineages is estimated. Some lineages have higher substitution rate than others; the question is why? Is it possible to find drivers/predictors for the variation in substitution rate among the selected species data? Evaluating a large number of morphological, ecological and life-history traits increases the chance of finding associations.

The study lacks hypotheses, which may be considered a problem in this kind of parameter rich correlative studies as any result may be open for quite speculative interpretations. Several correlations between the four genetic metrics (dS, dN, omega, inter-genetic) and the species and chromosomal attributes are detected (including with clutch size, beak size and chromosome type).

Response. The second and third paragraphs of the Introduction refer to the major exising hypotheses of the dominant drivers of molecular rates. These paragraphs have been revised so that they outline our hypotheses more clearly. For succinctness, we focus on introducing the major hypotheses as opposed to introducing those on each individual trait explored.

Originality and significance:

To my knowledge these data sets have not been combined and analysed in this way before.

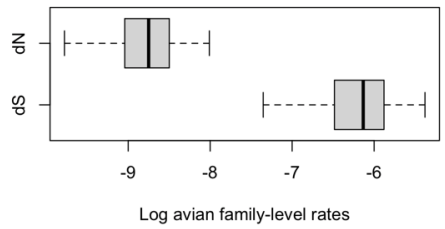
**Data & methodology: validity of approach, quality of data, quality of presentation;*
Appropriate use of statistics and treatment of uncertainties;
Suggested improvements: experiments, data for possible revision;
*References: appropriate credit to previous work?:**

Genetic data are from B10K (Fang et al.). This data set includes ≈370 species. Why are only 281 included here? Why were almost 100 species excluded? The authors must make a transparent statement of which were excluded and why. Maybe I missed this information?

Response. The original data set of Feng *et al.* includes multiple samples from some families. We aimed to test and interpret phenomena at the family level. By reducing the data set to the family level, our data set allows for more intuitive interpretation of results and reduces the problems posed by node-density effects, which can mislead family-level molecular rate estimates (doi:10.1111/j.1558-5646.2007.00188.x). This has now been clarified in the Methods section.

The genetic variation is measured as the rate of synonymous dS and non-synonymous dN in genic regions, and degree of substitutions in inter-genic regions. Also, the ratio between dN and dS, termed omega, is utilised – this measurement is supposed to be the most appropriate estimate of selection as other evolutionary processes such as genetic drift and mutation rate are expected to affect dN and dS in the same way. Note, however, that variation of the degree of selection and drift in different time periods might cause much more complex patterns than explained by these simple metrics at a specific time point. Overall, omega is the most appropriate metric to measure putative drivers of substitution rates in the genome unrelated to genetic drift and mutation rate. The section outlining the substitution rates (lines 88-95) contains errors I believe and needs to be revised and references given. "We tested the association between these traits and average family-level rates across four measures of molecular evolution: dN, dS, dN/dS (ω), and rates in intergenic (IG) regions. These measures have the potential to provide complementary insights into the evolutionary process: dN is influenced by the mutation rate, population size, and selection; dS is influenced by the mutation rate; and their ratio (ω) is influenced by selection and population size. Genome-wide variation in ω is expected to be driven by population size. Rates in intergenic regions of the genome are expected to reflect the mutation rate." Kimura's original derivations for dN and dS both contain 1/N, and thus N is not supposed to affect omega (i.e., contrary to the suggestion above). This probably affects many of the interpretations in this MS. Also rates in intergenic regions are affected by N.

Response. The impact of population size on d_N/d_S has been widely acknowledged and documented since the introduction of the nearly neutral theory, which revolves around the balance between selection and drift in different population sizes (Ohta 1973, doi:10.1038/246096a0; Ohta 1992, doi:10.1146/annurev.es.23.110192.001403; Yang and Bielawski 2000, doi:10.1016/S0169-5347(00)01994-7; among many others). There is also substantial evidence that d_S rates are nearly always neutral, including research reviewed by Kimura himself (Kimura and Ohta 1974, doi:10.1073/pnas.71.7.284). This is supported by our own results, whereby d_S rates are far greater than d_N rates, as shown in the figure below across the complete set of coding loci.



If synonymous rates are indeed representative of neutral evolution, then the ratio of d_N/d_S cancels out the mutation rate, leaving the effect of selection and population size. Another expression for d_N/d_S is $2Ns/(1-e^{-2Ns})$ where N is the effective population size and s is the selection coefficient, making it evident that d_N/d_S is affected by both of those parameters. Even in a case where s is large, a very small N can make nonsynonymous mutations behave in an effectively neutral manner.

The paragraph regarding the (lack of) results for omega is surprisingly vague. Basically, the logic and implications for most of the text at lines 158-171 are difficult to follow: "The lack of support for any association of genome-wide ω with body size or any other trait points to a limited role of variations in population sizes on avian molecular evolution. Population sizes are believed to have increased rapidly immediately after the K–Pg transition, during a period of newly available ecological niches. If this were the case, then we might also expect lower ω in smaller-sized avian taxa, consistent with the Lilliput effect, due to the more efficient removal of slightly deleterious nonsynonymous mutations24. One explanation for the lack of this signal in our data is that the effects of population-size fluctuations can be seen primarily on shallower timescales rather than in deep lineages where the signal has eroded. Alternatively, our finding of variation in ω across axes of variation and across chromosomes suggests that the impact of population size might be uneven across the genome (Stiller et al. submitted). Lastly, incomplete lineage sorting and reticulate evolution can also degrade the signal of ancestral population size39. Future examinations of these questions at finer taxonomic levels will be able to show a more nuanced picture, for instance by extending estimates of population size dynamics40 into deeper timescales." Why is the "K–Pg transition" interesting here? Why are "smaller sized taxa" discussed? What is the "Lilliput effect"? Why is "selection" suddenly discussed, and would not omega have been associated with some traits if selection was more prominent? What does "variation in ω across axes of variation" mean; variation explains variation? What is and why are "incomplete lineage sorting and reticulate evolution" discussed? I believe that the degree of incomplete lineage sorting is probably quite limited at the family level (most branches are >15 Myrs).

Response. We have extensively revised this paragraph to improve its accessibility to readers. In particular, we have aimed to improve clarity regarding our

expectation of ω in the face of a possible Lilliput effect. Our reference to axes of variation is also revised for clarity.

Line 195 and following lines. I suggest discussing the actual level of dN and omega as well, as low and high values are associated with very different processes.

Response. In our analyses, a principal component can be positively associated with rates in some branches but negatively in others. A principal component also considers the rates observed at all loci, so reporting single values of ω is neither straightforward nor intuitive. The result in line 195 focuses on the portion of all the variance absorbed by the first principal component, with contribution of the important lineages shown in Fig. 2. Importantly, the values of loci along a principal component are rates along new dimensionality that could be high for some branches and low for others, and therefore they are not directly interpretable.

The statement at line 205-206 needs a reference.

Response. We have added a reference to support this statement, as suggested.

Overall, you need to interpret PCs with caution (especially the ones explaining little of the variance).

Response. The method used focuses on the principal components that explain more variance than expected under permutation. This test exposes the principal components that provide significant amounts of information, even if they explain a small amount of the overall variance in the data. This can occur in principal components in which small subsets of loci are strongly explaining the rates in small subsets of lineages. In our data set, permutation tests revealed that only a small portion of principal components (<5%) explained significant amounts of variation in rates.

Line 208: To understand lineage specific patterns, it would be important to evaluate whether the multi-species alignment underlying these analyses is equally robust for different parts of the phylogeny. You need to discuss and clarify for the reader whether or not such potential issues might have affected your estimates of substitution rates.

Response. The genome-wide statistical support for the species tree is very strong across nearly all branches, as shown in Stiller *et al.* (2024). We have now expanded our discussion of high gene-tree incongruence in some of the important branches, and its consequences for the results of our study. Briefly, the main consequence of low support is that we cannot rule out that for many genes the phenomena we

observe have occurred in nearby branches that do not exist in our species tree, such that our result is a kind of average across the genome.

Figure 3. I'm not sure this figure shows "mean chromosome rate"? (3A) I assume positive values mean higher than estimated with permutation.

Response. As the referee points out, the data shown are not mean chromosome rates but rather the distance from the permutation distribution (z-scores). This means that high values are not high rates, but higher than expected under permutations of the whole data set to samples of the given chromosome size. This is now more explicit in the figure and caption.

It is surprising to see dN>dS, and positive omega, for a large chromosome like 7.

Response. d_N was never greater than d_S in our results. The reviewer might have misinterpreted the z-scores as absolute rates, when they are actually relative to other rates of the same type, and corrected by their variance across the data. Therefore, high d_N z-scores indicate these rates are higher than other d_N values, relative to the variance across these types of data. This z-score can therefore be high even if all d_N rates are all far lower than d_S rates. We have revised the explanation of our permutation in the figure and caption.

Also, this is the only chromosome with a large omega value, right? Could this be due to alignment errors, or translocations etc., in some part of the phylogeny.

Response. The high ω obseverd in chromosome 7 is due to reduced d_S. The chromosome has one of the most conserved syntenies across deep stretches of time (Schmid et al. 2007, doi: 10.1159/000056772), and this is likely to be associated with the low mutation pressure observed.

Note that chromosomal synteny is not conserved among all birds (there is variation in chromosome numbers, there are fused and split chromosomes, etc.).

Response. This study focuses on evolution at the nucleotide level, which is strongly grounded in theory. We now make it clear in our discussion that other measures of genomic evolutionary rates must be studied and modelled in future work, and we give multiple examples.

(3B) I think it is difficult to understand what all these PCs are telling (16 PCs (!) for intergenic are presented; e.g. what does PC 11 tell us?).

Response. The number of significant principal components is a signal of the complexity in rates. The greater the number of these components, the greater the

number of subsets of loci and taxa that explain significant amounts of variation. Therefore, each of these is potentially interesting. Some might expose groups of genes that have dominated the evolution in subsets of lineages, for example due to important ecological adaptations. Finding explanations for each of the signals is out of the scope of this study.

Line 257-260. "This finding on ω rates is consistent with widespread selective constraints on microchromosomes immediately after the K–Pg transition. This result is also consistent with the critical role that has been proposed for microchromosomes in harbouring housekeeping genes." You need to explain how these conclusions arise, and how they are compatible with each other and previous results and statements for the K–Pg transition (this is now unclear).

Response. In order to improve clarity, this paragraph has been revised and the argument explained differently altogether. We are now more explicit about our evidence revealing early members of Neoaves and passerines as having major molecular rate shifts in microchromosomes.

Lines 263-265. "The overrepresentation of microchromosomes in the evolutionary rate shifts along early neoavian branches suggests that they have had a greater influence on avian evolution than macrochromosomes." This is a very simplified conclusion and interpretation (possibly based only on visual screening?) of the results. Note that, in Figure 3A, chromosome 7 is the only chromosome with deviating omega – not the microchromosomes. In figure 3B some but not all microchromosomes show deviation signals for omega, and chromosome Z is also deviating, and only for some (few) PCs. Some of the chromosomes of intermediate size are showing similar patterns as the microchromosomes.

Response. Our interpretation is based on statistical tests across chromosomes. The results in Fig 3a show that chromosomes 1, 2, 3, 5, 6, and 7, are the only ones that are not significantly overrepresented across any principal components. This is evidence of a minimal role of these chromosomes in driving shifts in molecular rates. Instead, the first four principal components consistently highlight microchromosomes as more important than macrochromosomes in explaining variation in evolutionary rates.

Lines 271-274. "Since our results of molecular rates focus on evolution at the nucleotide level, they suggest that even under high mutational pressure synteny can be conserved across deep timescales in microchromosomes, reflecting a balance between high nucleotide mutation rates even under severe selective constraints." What is "they" referring to?

Response. We have rephrased this statement for clarity.

"…high mutational pressure synteny can be conserved…" I think you have the data to find support this statement (or not) in your alignment. You need to make it clear why high mutational rate is suggested (omega tells primarily about selection, as mutation rate is thought to affect dS and dN similarly).

Response. This stentence has been rewritten for clarity. Although our study does not explicitly examine synteny, we have revised Fig. 3 to provide better context for molecular rates across chromosomes.

**Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions;*
*Conclusions: robustness, validity, reliability:**

Line 276 and forward. How was "the most prominent branches driving rate variation" defined?

Response. For each significant principal component, these are the branches that explain significantly more variation than expected under permutation. This has been rephrased for clarity.

Also, there are so many options for gene enrichment analysis in this data set – the four variables (ds,… etc) and PCs (up to 16 for intergenic). This makes the results of this whole exercise quite difficult to evaluate. Any genome-wide genetic signal will show some genes being more associated with the signal than other genes – thus, enriched pathways can always be presented and suggested, and are not particularly strong evidence for processes in correlative studies, although it may look like there is a robust molecular basis (as e.g. implicated by the results presented in figure 4, and in the Abstract).

Response. Our analyses were performed while considering the possible pitfalls of gene set enrichment analysis, as outlined in the relevant literature (doi:10.1371/journal.pcbi.1009935). In addition to suitable correction of *p*-values and stringent thresholds, we find it extremely unlikely that our combination of results has occurred by chance. For instance, studies often find large numbers of enriched terms, while we consistently find the same small number of terms being enriched across several principal components. In addition, many components did not show any term enrichment. This is expected biologically if the sets of genes do not represent any specific known term, but a mixture, and indicates that our analyses are correctly filtering out misleading terms. Our focus on the third principal component of ω is also consistent with the substantial clustering of genes in higher chromosomes in this component (Fig. 3b). Lastly, it is remarkable that we almost exclusively find pathways that make up the ribonuclear machinery for

regulation. This seems contrary to a scenario with pathways that contain large numbers of genes and are unrelated to each other, such as those related to the cell cycle, endocytosis, and actin cytoskeleton, among others.

A potential main concern with this paper is the use of deep lineage genetic data and shallow data for the morphological, ecological and life-history data. Or in fact it is not clear when reading the main body of the MS whether the morphological, ecological and life-history data are from the species between included in the analyses or whether these data are reflecting properties of the clades? This would have been more appropriate as the genetic data (on the family level) likely reflect processes taking place long before the actual species appeared. This might be why the authors find most (putative) signs of selection at the terminal branches (which are in fact representing several million years of evolution; probably in most cases >15 Myrs).

Response. The manuscript has been revised to clarify that the rate estimates and traits reflect values at the clade level, taking their mean across families, rather than focusing on the sampled species. Specifically, rates summarize the molecular change along a whole family branch, so they represent the species sampled plus all ancestral lineages leading to the earlier split (i.e., changes including ancestors within the family). For completeness, we have now also included supporting results using trait data at the species level. As expected, these show far lower R^2 values in regression analyses explaining molecular rates, so they are not presented as the main results.

UNIVERSITY OF COPENHAGEN
DEPARTMENT OF PUBLIC HEALTH

Senior Editor, *Nature*,

26 OCT 2024

David A. Duchêne
Associate Professor
Department of Public Health
Øster Farimagsgade 5
1353 København K
+45 35 32 10 84
david.duchene@sund.ku.dk

Please find below our responses to reviewers for the second revision of our manuscript “Drivers of avian genomic change revealed by evolutionary rate decomposition”. We provide detailed explanations for the extended changes made in this version, as recommended.

Yours sincerely,

David A. Duchêne and Simon Y.W. Ho, on behalf of all co-authors

Referees' comments:

Referee #1 (Remarks to the Author):
Review for: Drivers of avian genomic change revealed by evolutionary rate decomposition

This is the second time I have reviewed this paper, and while it has improved in some respects, I still have concerns that prevent me from recommending publication in its current form.

Below, I provide some responses to the responses that the authors have offered to my original review, and then I follow with additional comments/questions going line by line.

That said, the most significant concern I have with this manuscript is that many of the regression results emphasized by the authors are "statistically significant" but with low effect sizes (my guess; see my line-by-line comments below). I have not been able to figure out where to access 'Supplementary File 3,' (Tables of phylogenetic regression data and results), so it's not clear if further information reported there would help me evaluate these aspects of the work.

Response - This is an important point that we have considered very carefully in our revision. To quantify uncertainty in parameter estimates naturally, we have now opted to cast our whole regression analysis onto a Bayesian framework. We have also added random forest analyses as suggested in other comments. We acknowledge that there are many benefits from these approaches to our type of data, and we have invited Professor Samir Bhatt to contribute his expertise to our study. The uncertainties in our coefficients are now reported in detail in the form of credible intervals. Supplementary code, result files, and additional figures have been updated and are available at github.com/duchene/avian_family_rates.

In regression analyses, arbitrarily low p-values can be achieved (even considering correction for false discovery rate) with a dataset of this magnitude simply due to extreme sensitivity (very small effects can be detected). On one hand, this is a potential benefit of a dataset like that employed here. However, on the other hand, if many of the results represent minor effects (even if statistically significant), the significance of the overall conclusions is weakened.

For example, I am concerned by the lack of reporting of effect sizes for every test I can see in this manuscript — p-values are reported with R^2 values, but no test statistics or relevant model coefficients (effect sizes) are reported in the main text. I wonder if this is because the effect sizes are small? More

transparency is needed to be able to evaluate these patterns. Further, the R^2 value estimated by phylolm is likely not interpretable in the same way that a standard coefficient of determination is, as the data are hierarchically structured (a violation of standard linear regression); see this paper ('R^2s for Correlated Data: Phylogenetic Models, LMMs, and GLMMs': https://doi.org/10.1093/sysbio/syy060). These comments reflect a general concern about the regression approaches employed here.

An alternative approach I might consider could be something from the field of machine learning, like random forests regression (with an adjustment for phylogenetic signal), which may provide a simpler assessment of how each of the 20+ predictor variables contribute to the overall predictive accuracy of a regression model (e.g. for predicting a given molecular rate). Given the scale of the data here, this might be very powerful. For that type of approach, as an alternative to model coefficients, one can easily extract individual variable importance scores, as well as an overall understanding of how much predictive information is contained within the predictors (e.g., AUC). These machine learning approaches also make essentially no assumptions about the form of the underlying data, and so may be well suited for the task here; no model selection or predictor selection is required (just give it everything and let the black box sort it out). This is an entirely different approach than what has been employed here, so I would leave it to the editor to decide if some exploration like this should be attempted. Nonetheless, I am so far not persuaded by most of the regression analyses that are reported, as they seem to imply small, though statistically significant, effect sizes.

Response - Figure 1 now reports the estimates of the multiple Bayesian regression coefficients and credible intervals across variables, as well as importance scores from our random forest analyses. In the text, we now refer to the results using Bayesian credible intervals, and interpret them in terms of the combined signal across our analyses. The supplementary material reports the same results across various treatments of the data (family- and species-level traits), and includes the results from the random forest and frequentist counterparts of our regression analyses.

#######

Below, I provide additional responses to responses that the authors provided in response to my initial review. For the five responses below (noted numerically), I have copied the author's response and provided an additional comment.

1)

Author response:

Response. Residual variation in regression was indeed modelled using Brownian motion, with the amount of phylogenetic signal in the residual error modelled using the variance rate σ2 (this is equivalent to using λ, and the two parameters are non- identifiable; Ho and Ane 2014, doi:10.1093/sysbio/syu005). Values of σ2 were higher in regressions involving genome-wide molecular rates of dN, dS, and intergenic regions than in regressions of principal component loadings. This is intuitive, showing stronger phylogenetic signal in genome-wide rates than in the rates decomposed into subsets of lineages and genes. For interested readers, the data repository now points to the estimates of σ2.

Reviewer response:

There seems to be some confusion here. σ2 and λ are not equivalent parameters and are not used in the same context in phylogenetic comparative methods — the former is typically interpreted as the rate of character evolution, while λ is an estimate of a branch-length transformation parameter sometimes interpreted in the context of phylogenetic signal. In any case, values of σ2 are not appropriately interpreted as more or less 'phylogenetic signal;' these are different things. I suggest the authors may want to consult Luke Harmon's textbook on these topics.

The paper that the authors refer to does not include the word 'identifiable,' nor would it make sense to try to estimate these parameters at the same time, so I do not follow the interpretation offered here. σ2 is not a measure of phylogenetic signal, it is a parameter that estimates the rate of character diffusion or variance. In the context of regression, σ2 is used to account for variation in the residuals, assuming a pure BM process. If instead λ transformation is estimated/used, a pGLS will try to account for a mixture of BM and non-BM processes in the model residuals (this is described in the paper the authors reference above in the section 'Modified Brownian Motion Models.'

Response - We thank the reviewer for pointing out the distinction between character diffusion rate/variance (σ2) and phylogenetic signal (Pagel's λ). We acknowledge that the two parameters are formulated differently; yet, both relate to the variance in the response such that they are non-identifiable. This is described further in the phylolm R package documentation (https://cran.r-project.org/web/packages/phylolm/phylolm.pdf). We now

focus our interpretations on the results that are consistent across our Bayesian, random forest, and frequentist analyses. Bayesian regression analyses include the phylogeny as a random effect with a shrinkage term. As such, the impact of phylogeny is also an inferred parameter. The inferred parameter naturally incorporates the uncertainty in the impact of relatedness across all other effects.

2)

Author response:

Response. Regression analyses have been added using the species-specific data rather than family averages, and these results are available in the supplementary material. We find that species-level regressions have far lower R2 values than family averages. This is expected given that molecular rate estimates correspond to the full history of molecular evolution along the lineages from each sampled individual back to the common ancestor shared with its sister family. In this sense, molecular estimates are not specific to the sampled species but more broadly represent an average rate across the history of the family. As such, the family-level trait average is far more likely than the species trait to be associated with our molecular rate estimates. This 'whole lineage' aspect of our molecular rate estimates is now explained in the text, as well as our comparison with species-level data.

Reviewer response:

I still do not fully agree with the authors' logic here. The "family average" is a problematic representation of the underlying trait data and does not truly correspond with the genetic samples in the dataset. As I suggested in my prior response, running the regressions with resampled trait values from within the family is perhaps more palatable but, even still, sub-optimal. An alternative approach emphasizing "family averages" could be to estimate an ancestral state for each family (a phylogenetic 'average'), prune the terminal lineages to the family MRCA age, and then use the estimated ancestral state values; this would attempt to avoid the sources of bias I am concerned about. In any case, a 'phylogenetic' average would be required to account for what the authors describe in their response, and not a standard mean.

Also, see my comment above about the interpretability of the R^2 value from phylolm; I suspect that a different approach is required if the goal of each individual regression analysis is to assess how much variance in the data is explained by the model (e.g., R^2pred, as noted in the paper by Ives).

While it is true that the molecular rate estimates as presented will reflect the average rate across the history of a branch (+/- some residual variation due to taxon sampling), it is also true that a biased taxon sample (in this case, a limited representation within families) is expected to lead to molecular (or any character) rate underestimates because the underlying substitution model rate matrix cannot 'see' the missing lineages (e.g., missing substitutions). It is not possible to know how much bias is introduced by using family-level trait averages without measuring or simulating it; I leave it to the editor to decide if further investigation is required, but I am not fully convinced by the authors' justification.

Response - These are all reasonable points that have been discussed previously in the literature of molecular evolutionary rate analysis. The use of a family-level average has been widely regarded as the best approach for large-scale analyses of molecular rates (e.g., doi.org/10.1038/ncomms2836), and is far preferable to single-species or ancestral state estimates. Single species will rarely provide a good summary of the family-level trait (as done by the molecular rate estimate). Ancestral state estimates will have error compounded over a series of additional models, all of which are also likely to suffer some form of misspecification. This includes making further assumptions about molecular substitutions, clocks, diversification, and trait evolution, additionally coming from incomplete species-level data. Since there is no single widely accepted phylogenetic estimate covering all avian taxa, the excess error in any ancestral state estimates would only produce a questionable improvement, if any, compared with the family-level mean. We provide the data and results for both family- and species-level analyses, but interpret results using family-level data.

Molecular rate estimates can indeed be influenced by punctuated changes upon speciation (doi.org/10.1126/science.1129647). Similarly, the inclusion of additional samples alone can lead to the identification of additional substitutions (doi.org/10.1111/j.1558-5646.2007.00188.x). Separating these two sources of signal is not feasible, and existing solutions are not convincing (doi.org/10.1126/science.1083202). Our approach provides a widely accepted balance where nodes are not present in the calculation of substitution rates (doi.org/10.1016/j.tree.2010.06.007), and instances with extremely limited signal and high uncertainty are not included (doi.org/10.1016/j.jtbi.2007.12.015). We have now added a note in the manuscript about the limitations to using family-level estimates.

3)

Author response:

Response. Mass-corrected variables are used in the figure for visualization purposes, but this was not the case in regression analyses. Instead, body mass was included as a covariate in every regression analysis. This addresses its joint impact with other variables on molecular rates. For visualization purposes, variables were mass-corrected in cases with a strong association with body mass, such as in the case of generation length ($p < 0.001$, $R2 = 0.67$), beak depth ($p < 0.001$, $R2 = 0.65$), and brain size ($p < 0.001$, $R2 = 0.89$). In these cases, body mass has a strong impact on visualization of the ancestral states. In contrast, the correction has limited impact on the visualization of clutch size, which is not predicted strongly by body mass ($p = 0.02$, $R2 = 0.03$). The figure and caption have been revised to clarify these points.

Reviewer response:

The author's response does not address my original comment: What is a 'mass-corrected variable' in this context? The authors' response does not actually explain what they mean here. Does this mean that the data points in the figures are mass-corrected residuals from a regression model? Are they normalized by mass (e.g. divided by mass?)? Further, it is not clear what it means to use 'mass-corrected variables "for visualization purposes" — does this mean that the data points depicted in the figures are not the same as the data points used in the regression analyses?

Response - The correction of plotted variables by body mass was performed via regression residuals, and the code for this is available in the supplementary material at github.com/duchene/avian_family_rates. The correction was performed for visualisation in figures, since using the raw variables would mean that the colour primarily reflects variation in body mass, rather than the variable, as described in our previous response. In regression analyses, the relationship between the variables and body mass was addressed by adding body mass as a covariate.

4)

Author response:

Response. This is a particularly useful comment and has led us to implement a more sophisticated approach to multiple regression. We now consider every possible combination of variables in our multiple-regression models, and weight models by their AICc for model averaging. We report the model-average results across weighted models. The results are consistent with

those of previous analyses with the single exception being the results for dN rates. In this case no variable consistently emerges as a strong predictor across all analyses. Traits that are found to be significant predictors in subsets of analyses are discussed further in the text, including beak depth, brain mass, and tarsus length. Some analyses that show such traits are those where all palaeognath taxa are included, or multiple-regression models based on single-variable analyses.

Reviewer response:

The model averaging procedure is not fully described in the methods; this is required to better understand what is happening here. For example, are the authors weighting the model coefficients by the alternative model weights? It's not clear.

Response - The results are now based on a single full model, using a combination of Bayesian regression, random forest, and frequentist regression, and we hope that this is more intuitive. Our interpretations focus on the results that are congruent between the full model in Bayesian and frequentist regressions. We have also extended our descriptions of the methods for clarity. Importantly, we now focus on analyses that include all variables in the same model, such that we avoid variable selection via *a priori* simple regressions or model averaging.

5)

Author response:

Response. The recovery of deep branches in this species tree is likely to be affected by several forms of inference error. At this deep scale, model-misspecification error arising from true gene-tree discordance (i.e., incomplete lineage sorting) will be outweighed by gene-tree inference error. This is demonstrated by very low branch supports in individual genes at deep branches. Importantly, tree inference error is greater at deeper branches, while the error due to incomplete lineage sorting is similar across the tree even with fluctuations in population size. An excellent description of these relative amounts of error is given by Bryant and Hahn (2020, section 4.2, doi: hal-02535651).

Reviewer response:

This seems to be an argument in favor of excluding genes with low branch support on deep edges, as this would imply that the error in estimating branch lengths near those deep edges would also be high (or that there is very little signal of branch length information).

Response - Genes with branches that are discordant with the species tree, or with low branch supports, are all likely to contain far more phylogenetic error compared with our species-tree estimate. This is why cases of excessively short and long branches were filtered out. Average rates across all genes are also likely to be more reliable than estimates in single genes. Similarly, patterns observed across large numbers of genes are more likely to show an interpretable signal than in single genes (e.g., results from rate decomposition and functional term enrichment).

#####

Line-by-line comments:

Abstract: —

Line 18 ' we find that the dominant predictors' — rephrase to say, 'of the traits we evaluated' ...

Response - This sentence has been rephrased, but in light of our latest results we have chosen to maintain similar wording to our previous version.

Line 107 —

R2 is not a particularly useful metric without also knowing the model coefficients (particularly for multiple regression)—e.g., the effect sizes. A large R2 and a low effect size do not suggest a robust relationship. Also, see my prior comment about R2 for correlated data.

Response - Figure 1 now includes Bayesian coefficient estimates and credible intervals across variables and analysis runs.

Line 110— same comment as above — what is the standardized effect size of the noted predictor in the optimal model?

Response - We hope that the uncertainty in coefficients as shown in Bayesian credible intervals provide a more transparent measure of effect size.

Line 130—

Every time statistical relationships are reported (when a significant 'predictor' is being emphasized), I think it is important to report the estimated standardized effect size (e.g., the relevant model coefficients), the p-value, and optionally the appropriate coefficient of determination R2 (as suggested by Ives). In my opinion, relationships that have a low effect size (model coefficient) do not suggest a strong effect, even if the p-value is low.

Response - Coefficient estimates and uncertainties are now provided in detail in Figure 1 to depict effect sizes across model parameters.

Line 153-158

Can the authors clarify—lines 157-58—are the models testing body size set up as 'y ~ body size + covariates' — e.g., directly testing the effect of body size while including other covariates in the model? If so, what is the standardized effect size for body size, p-value, etc., as above?

Response - The standardised effect size for all parameters is now shown upfront in Figure 1a.

Line 166
The suggested 'lilliput effect' is a hypothesis and not firmly established (e.g., through fossil evidence). Maybe should qualify it a bit more cautiously here.

Response - The text is now more cautious when describing the Lilliput effect.

Line 170-172

Other studies seem to be able to detect population size fluctuation with whole genome data (e.g., Stiller et al. 2024, and Houde et al. 2020). I suspect one reason it may not be detected here could be because the data are not analyzed in a coalescent framework. I have no strong opinion that this is a better way to analyze the data; however, now that it has been reported in two independent studies, the fact that it is not observed here suggests to me that the approaches used by the authors have introduced a source of bias/noise into the results (perhaps due to fixing the gene tree topologies to the species tree topology).

Response - This metric is only partially affected by population size, and in fact is traditionally used for measuring changes in selective constraints, with population size being a factor driving possible bias. The text is now more explicit in describing that it is impossible to distinguish the relative contribution of selection versus population size on the results.

Line 182

What is a 'single gene rate averaged across lineages' and how is it computed?

Response - This is the overall evolutionary rate of a gene. This is taken as the unweighted mean rate across all branches, such that it is averaged across

all taxa included in the data across their evolutionary history. This explanation is now extended in the manuscript.

Line 185-187

Again, what are the effect sizes? — they are not reported in the main text. I suspect the relationship between GC content and rates may be due to model misspecification in genes with high GC content. This is now well documented in the literature — e.g., Stiller et al. 2024 exclude loci on this basis. Loci with higher GC content are more likely to violate model assumptions of stationarity, which could cause apparent rate variation (e.g. more rate variation across the entire phylogeny could look like a higher average rate). Non-homogeneous substitution models may be required to assess this more robustly.

Response - Figure 1 has also been extended to include the standardised coefficient estimates with credible intervals for the regression analyses on gene rates. Therefore, we are now more explicit in reporting the effect size and its uncertainty, as suggested.

Line 194-196

This supports what I suggest above regarding model misspecification.

Response - In order to examine this question further, we repeated our regression analyses by adding model mis-specification statistics per gene. These include non-parametric test statistics for overall model fit (doi.org/10.1007/BF00166252) and compositional heterogeneity (doi.org/10.1080/10635150490445779), as well three parametric p-value results for non-homogeneity in the molecular evolutionary process (doi.org/10.1093/gbe/evz193). All regression results were nearly identical after inclusion of any of these variables, such that any possible link between model mis-specification and GC content or molecular rates is not driving our results. Our data sets now include all of these metrics of model adequacy for each gene (github.com/duchene/avian_family_rates).

Figure 1

Again, it is challenging to interpret the models without clearer reporting of the effect sizes. Line 202—'mass corrected' is again not clearly explained. Is this model residuals? Is this generation length/mass?

Response - This case also involves regression residuals, and this treatment is only relevant for figures. The caption of Fig. 1 has an extended description of the data used for colour scales.

Line 220

How many components is < 5%? What is the effect size/test statistic? — just reporting the p-value does not seem sufficient.

Response - We have revised the reporting of these results to be more upfront about the number of significant components. The large number of test statistics for principal components are available in the Supplementary Files.

Line 224

The terminal branches popping out here are suspect to me and suggestive of some kind of artifact/bias. My guess is that it may be due to sequencing error, which I would expect to induce extensions of terminal branches.

Response - Far from being suspect, this result is fully consistent with theoretical expectations under molecular clock theory (doi.org/10.1093/oso/9780195068832.001.0001) and with our previous work on a large number of data sets (doi.org/10.1093/molbev/msz291). It reflects the fact that the dominant form of variation is across loci. The result is extremely unlikely to be due to sequencing error, which is negligible in this data set (doi.org/10.1038/s41586-020-2873-9).

Lines 236-245

I am not sure we can rule out that these patterns are due to rate estimation errors at nodes with high levels of incomplete lineage sorting.

Response - Inferred substitutions can become misplaced due to gene tree incongruence (referred to by previous authors as 'SPILS'), and these can be incorrectly placed on more recent adjacent branches (doi.org/10.1093/sysbio/syw018). In the text, we are cautious not to overstate the presence of an excess of substitutions on a specific branch, such that these might occur at adjacent branches in genes with different tree topologies.

Line 285

If this is a permutation test, what is the test statistic and its value/interpretation? It should be reported here.

Response - The tests on principal components are based on two test statistics that summarise the signal of eigenvalues (ψ and φ; doi.org/10.1111/evo.13835), and tests on lineage loadings are based on

loading values themselves (Supplementary Data Fig. 1). This is now described in additional detail in the text.

Line 297-299:

Really cool result!

Line 303-305 — If true, this is the most compelling result of the paper, in my view.

Response - We have given high visibility to this result, and hope that other results are also compelling given our latest round of analyses.

Line 324— again, what is the test statistic etc.

Response - The *p*-value in gene overrepresentation analysis (doi.org/10.1093/bioinformatics/bth456) is equivalent to a one-sided Fisher's exact test, such that it does not involve a classical test statistic other than the probability itself. The test focuses on the probability that a given set of genes (associated with a given gene ontology or KEGG term) is overrepresented in any other given set. We have extended the description of this test in the main text and methods sections.

Line 330, what is the effect size? I suspect it is extremely low, given the R2. In this case, the significant p-value is possibly a consequence of the massive sample size, suggesting very high sensitivity to extremely minor effects. I see that the approach with model averaging (not clearly explained) still includes tarsus length, which is encouraging; however, the effect size still seems to be extremely low. I suggest reframing in terms of model selection.

Response - The effect sizes with Bayesian credible intervals for this analysis are reported in the supplementary data and figures (Supplementary File 4). These show that most parameter estimates have credible intervals that are broad and overlap zero, such that any results that do deviate from zero are meaningful. Intervals are also nearly identical between our Bayesian and frequentist regressions.

Line 334

The authors refer to different tarsus 'optima' — but it does not appear that they have fit models of stabilizing selection (e.g., OU models) that would allow them to make this statement in a quantitative way.

Response - We believe that our statement does not require quantification of tarsus optima, since this is intended to be a qualitative reference to the

diversity in modern avian taxa, rather than about parameter optimisation in statistics. The text has been modified to avoid any potential confusion.

Figure 4a

The results discussed here appear to show a very marginal effect. As I noted elsewhere, the effect size is not clearly reported.

Response - In this revision, the effect size is made available in tables and figures (Supplementary File 4). Credible intervals and our interpretation of results only if congruent across our Bayesian and frequentist analyses provide a strong assurance that any meaningful result is worthy of interpretation.

Line 400

I am not convinced that the paper has demonstrated this; it seems like the study shows a variety of significant correlations with low effect sizes.

Response - Our casting of regression analyses provides several key improvements from traditional effect sizes, confidently placing our results as meaningful descriptions of the links between trait variables and decomposed molecular rates. Credible intervals are the directly sampled posterior probability that the effect size is meaningful. This not only provides a measure of uncertainty, but one with an intuitive probabilistic interpretation (the posterior probability of the model given the data). We hope that the reviewer agrees that our new analyses provide maximal robustness to our conclusions.

Line 468

These results look nice — the authors should report the effect sizes for each regression they report a p-value for. Including those values in supplemental data is not sufficient for transparency (though I have not been able to access these data tables). In any case, 'mass-corrected' is not clearly explained here, nor is it explained in the response to my prior comment.

Response - Figure 1 now includes the standardised effect sizes for these results, with posterior credible intervals. We now refer the reader to that figure.

Line 498:

Did the authors test for node density effects? It might be valuable to try to account for this directly, especially since the rate estimates are not coming from molecular clock models.

Response - Rates in this study are estimated only along single branches, such that each estimate is unaffected by nodes. There could be overall biases if substitution models are overestimating terminal branch lengths (doi.org/10.1080/10635150500354647), or from molecular change being linked with speciation events (doi.org/10.3389/fgene.2017.00012). However, the non-identifiability of rates and times makes it impossible to test these possible forms of bias, and it is widely acknowledged that estimates at single branches give the best chance of reliable comparative analysis of molecular rates (doi.org/10.1016/j.tree.2010.06.007).

Line 506-509

I remain concerned about forcing the tree topologies and the biases that this is expected to induce in rate estimates.

Response – Where the topology of the gene tree topology actually differs from that of the species tree, nucleotide substitutions can be incorrectly placed on more recent adjacent branches. This bias has been described in the past using simulations (doi.org/10.1093/sysbio/syw018). We are explicit about this possibility in the text; however, estimation error in the tree topology is likely to overwhelm any discordance due to incomplete lineage sorting, so we lack certainty about the true gene tree topology. Our approach of forcing a strongly supported topology, followed by careful data filtering, provides a balance between maximising information and minimising extreme signals that are likely to be misleading.

Line 520-526

The concern here is not that the process may have occurred on a nearby branch — it is that the inference of a rate change might be entirely artifactual. I remain concerned about the choice to emphasize results derived from analyses of forced tree topologies.

Response - In cases where a gene does not have the signal for a given branch, the length will be close to zero and excluded from our analyses. Gene-tree estimates can be highly inaccurate at such deep timescales, so they do not provide a reliable inference of the gene history. Conversely, our methods are designed to strike a balance across various forms of uncertainty: we exclude extremely uncertain branch lengths and use a realistic (species tree) topology that is likely to have strongly shaped the branching structure of the gene trees.

Line 538

"Continuous variables were verified to follow logarithmic scaling." — How did the authors do this? Typically, we log transform because it helps data conform to model assumptions, or because we believe that proportional scaling is more relevant than linear scaling. The quoted sentence means something entirely different.

Response - Most of the traits explored here scale logarithmically across taxa. We visually explored that this was the case, such that log-transformation led to better adjustment to gaussian model assumptions.

Line 541-547

I do not agree that a family average is always a good proxy of a phylogenetic average, which is what the authors are trying to estimate. These values may be very different; see below.

Response - The family average is likely to be the best treatment of the raw data, since other forms of averaging such as ancestral state estimation will actually compound several types of modelling error, including those in the inferred topology, divergence times, and trait evolution.

Line 548-550

I do not agree that models with higher R2 values using family averages necessarily support this inference. Perhaps this would be true if this were based on effect sizes, but it is not clear if that is the case.

Response - As explained above, using family-level means is well justified and standard practice in the field, but we have included data and results at the species level in the supplementary material.

Line 551-557

I am still not sure that this approach is statistically justifiable—the model averaging approach also does not directly address my prior concern. The potential causal interactions among the examined traits mean that effects may only appear when multiple variables are included or excluded in particular models (a causal modeling framework like phylogenetic path analysis would be required to disentangle some of these interactions). Therefore, I am still wary of the approach that restricts multiple regression to only consider traits significant in 1:1 analysis.

Response - The revised regression analyses under a Bayesian framework focus on the complete set of variables as covariates, such that only the full model is evaluated.

Line 558-561

Can the authors provide more detail about the logic of model averaging in this context? My prior comments about this did not request model averaging—they were about justifying the inclusion or exclusion of particular predictors on the basis of relative model fit.

Response - Our methods now emphasise the results from Bayesian regression, and we include analogous random forest and frequentist regression analyses for completeness. In brief, model averaging uses an AIC weighted average of coefficient estimates across models. Results focus on full models only, rather than model averaging, acknowledging that some variable combinations might be overly influential.

Line 565

Perhaps replace "Drivers" with "Correlations"

Response - This term has been revised for accuracy.

Line 573:

The authors again indicate model-averaging, but few details are provided as to how this procedure was carried out. I see that the description is included in line 990 about the MuMIn package but no detail is provided about what is being averaged.

Response - Our emphasis now lies on the full model as tested in Bayesian phylogenetic regression, random forests, and frequentist regressions. We interpret results that are consistent across all three analysis frameworks.

Line 590-594

It is still not clear to me that this procedure is statistically valid; the issue with mapping gene tree rates to species tree rates is not only about rates on a focal branch vs. a neighboring branch but rather about substitutions inferred on a focal branch that does not actually exist.

Response - Since the gene trees have large amounts of topological error (which compounds through time, while error due to ILS does not; hal.science/hal-02535651v1), the species-tree topology inferred from

genome-scale data is a reasonable estimate of the branching process for any given gene.

Referee #2 (Remarks to the Author):

I think the manuscript drastically improved and now incoporates GC content as one of the most important covariates of molecular rates. I happy how the authors addressed my concerns, and I think in the current form the paper will set a bar on comparative genomic approaches.

Response - We thank this reviewer for the positive remarks.

Referee #2 (Remarks on code availability):

I reviewed some of the code

I noticed that one of the scripts refers to a local library/file source("~/Dropbox/Research/code/brsPerTslice.R")

I think these should be avoided. The code should be able to run on its own.

Response - This has been revised as suggested.

Referee #3 (Remarks to the Author):

This is a revised version of a manuscript I previously commented on. The authors have made substantial changes in the new version and have sufficiently addressed previous comments.

Response - We thank this reviewer for the positive remarks.

Responses to Reviewer 1

In this revised submission, the authors have graciously and thoughtfully responded to all of my prior comments, and I thank them for their efforts to improve the manuscript. The addition of Bayesian regression and random forests analyses is impressive and improves the statistical rigor of the manuscript. These techniques replace previously opaque model averaging, so this is in the right direction. Slightly more detailed reporting of the random forests analyses would be welcome (particularly with respect to model performance). The authors note that they assessed model performance with RMSE, but those details do not appear to be reported. Were there any other assessments of the random forests performance?

Response. This version includes the full set of root mean squared error (RMSE) diagnostics as supplementary information. This means that the diagnostics are fully available to readers.

While I do not agree with all of the provided explanations to my prior questions (principally those related to the emphasis of a fixed tree topology), this perhaps stems from differences of opinion, and I am happy to co-exist with differences of opinion. I am generally satisfied with the other responses offered by the authors, and the paper has been significantly improved through review. Given the brevity of their comments, the other two reviewers do not appear to have reviewed the revised manuscript carefully; considering the substantial changes in the latest revision, it may be important to request additional reviews from new reviewers. However, should the minor changes I suggest here be incorporated, I don't think I need to see the manuscript again (at the editor's discretion).

Response. We thank the reviewer for the overall positive assessment. In previous revisions we aimed to be clear about the possible drawbacks of assuming a fixed tree topology, so we hope that our article provides a well-balanced view on the results.

Here are a few minor other comments for the next and hopefully final revision:
Since I read the last revision, there is a new pre-print linking rates of genomic evolution to tarsus length in birds https://www.biorxiv.org/content/10.1101/2024.04.30.591925v1 -- the authors may want to cite this pre-print where appropriate (though extended discussion of a pre-print is probably not warranted).

Response. We thank the reviewer for drawing our attention to this work. We have revised our discussion of tarsus length to mention its likely wider impact on the genome, citing the preprint accordingly.

The paper could also be improved with an enhanced/expanded discussion that more thoroughly compares their new results to the results of prior studies (some recent), specifically in the conclusion section. This is important and should be the last issue that the authors need to address (in my opinion). For example, this sentence appears in the conclusion:
"These analyses show that the key drivers of genome evolution in birds include clutch size, generation length, and tarsus length – traits long considered important in avian life history and ecology – along with GC content in genes."
All of these factors have been previously investigated with respect to patterns of avian molecular evolution, and relevant works should be cited here (some are cited elsewhere in the paper). Any

appropriate commentary (as judged by the authors) can also be added to help readers place these results within the broader research context of these topics. For example, the recent work by Berv et al (2024) https://www.science.org/doi/10.1126/sciadv.adp0114 suggests a weak link between clutch size and the mode of molecular evolution (as distinct from (though related to) the rate) -- and it is interesting that clutch size also comes out as an important factor in this analysis of a much larger genomic dataset (among other interesting/contrasting patterns). As the authors are aware, many other papers have investigated various life history traits and patterns of molecular evolution, and the conclusion section could be better used to emphasize how these new analyses build on prior results and the work of other researchers (including, for example, these important works:https://doi.org/10.1093/molbev/msx236, https://www.science.org/doi/10.1126/science.aat7244, https://www.cell.com/current-biology/fulltext/S0960-9822(17)31081-3, which investigate similar questions using model-based approaches). It may be especially important to emphasize how these new results differ from any prior scientific consensus (for example, the proposed associations with body mass). I suggest that the authors carefully review their concluding statements in the context of prior work and add more detailed acknowledgments. I know that Nature has strict length limitations, but I would like to ask the editors to give the authors some grace in including additional contextual details that will improve the scholarship of the work.

Response. We have succinctly revised the sentence cited and the discussion and conclusions to describe existing work on the links between avian molecular change and traits. We hope that this more clearly acknowledges previous efforts, while emphasizing the novelty in our work.