****Recommender's comment (Bruce Rannala)**

**This is an interesting, well-written paper examining the relationship between gene expression levels and the strength of purifying selection (as measured by pN/pS) in two species of penguin. Some compelling patterns emerge suggesting that increased gene expression levels at a locus are associated with increased selection.

>>We would like to thank the Recommender for reading and commenting on the manuscript and for the positive consideration of our work.

**I have read the paper and the comments/concerns of the two referees and largely agree with their comments and suggestions for changes. However, I will leave it up to the authors whether they wish to follow reviewer Pyhäjärvi's suggestion to reorganize the materials and methods incorporating more content from the extended methods section.

>>We have addressed all of the referees' comments in the revised version of the manuscript. Below you can find our point-by-point response (>>, in blue) to each Recommender and Reviewers' comment (**, in black). After consideration, we decided to keep the format as it was originally (Extensive Material and Methods at the end of the manuscript) as we feel it helps to better streamline the messages of the main text. However, we moved the description of the hypotheses tested with our analyses in the main text (lines 73-85) from the Extended Methods section as suggested by reviewer Pyhäjärvi's suggestion.

**An additional concern I had that was not specifically mentioned by either reviewer is that the idea of deciding between population size versus gene expression as the "main driver" of purifying selection appears logically flawed. As the authors know, the strength of purifying selection should be proportional to Ns. Gene expression alters the phenotype and therefore changes s. With larger N (if one believes the classical theory) the expectation is that a change of s will have a proportionally larger effect, this does not mean that s is the main driver of selection. Trying to partition the effects of N and s only seems to make sense if their effects are additive, but they are in fact multiplicative.

>> We also agree that gene expression is only proportional to s, and not to Ns, so we are not certain we fully understand this concern. Perhaps we were not clear in our previous version of the manuscript.

>>The relationship "2Ns ~ selection efficiency" comes from the probability of fixation of a new mutation which is a combination of drift P(fix_d) = 1/2N and selection P(fix_s) = 2hs. Disregarding the dominance coefficient h, for selection to be effective |s| >> 1/N, that is, weak selection can be effective only if the population size is large or that strong selection is effective also when population size is small. It seems then to be legitimate to ask whether the (non)equivalence above is fulfilled as a consequence of large *s* or large *N*, even if they are multiplicative. For example, after finding different selection effects at a small fraction of X-linked genes, Andolfatto et al (2011) suggested that a surprisingly high proportion of mutations have such a high fitness effect (or selection coefficient) that they were not affected by the different population sizes of the two *Drosophila* species under investigation.

>>In the case of the biological systems we investigated in our study, the observed value of pN/pS cannot be produced in simulations using the highest value we tested for Ne (100,000 breeding individuals) which is biologically reasonable for these species (e.g., Emperor penguin census size is around 500,000 breeding pairs and Ne is estimated as 40-50,000 - Cristofari et al 2016).

>>On the other hand, according to classic evolutionary models, most of the mutations should either have no (i.e. neutral), or a very weak effect on fitness, which equates to a small

selection coefficient. In fact, the distribution of fitness effects (gamma with mean -0.013 and shape 0.19, Kim et al 2017), which is commonly used in simulations, is in line with this (quasi)neutral model.

>>Of course, with larger *N*, the expectation is that a change in *s* will have a proportionally larger effect. However, when observing the intense selection effect in the Emperor and King penguin dataset we can ask whether it is an unreasonably large population size (for this biological systems) or a higher than expected selection coefficient which is causing |s| to be much larger than 1/N. In other words, if *N* is kept fixed for all of the genes, what is the range of values of *s* which is needed to reproduce the proportion of purifying selection that we observe?

>>We have tried to clarify these objectives more concretely at the end of the second paragraph, also addressing a comment from reviewer Pyhäjärvi (lines 63-70).

**There is also the issue that the population size differences are unknown with only the ranked population sizes extrapolated from differences of diversity, Tajima's D, etc. Since the locus specific mutation rates for non-coding DNA should be similar between species why not estimate theta for each species and compare them to determine the proportional difference of effective population size? Does it make sense to try to examine the effect of differences of population size for the penguin species when only two species/populations are available and only the possible rank order of size difference is known?

>>We understand the concern of the Recommender as we are comparing only two population sizes. However, the simulation part of our work aimed at strengthening this weakness. To this scope, simulations were performed across three orders of magnitude of population sizes (from 1,000 to 100,000).

>>Moreover, in the revised version of the manuscript, we framed the comparison between the two population sizes within a theoretical model that predicts the strength of purifying selection acting on the protein. As now explained in the text (lines 50-55), “Assuming that proteins are selected for their conformational stability (*i.e.*, the protein is folded or not) or for protein–protein interaction (*i.e.*, the protein is bounded or not to other proteins), the intensity of purifying selection acting on the protein can be theoretically derived as a function of both gene expression and effective population size (Latrille & Lartillot 2021), but so far the predictions of these models have not been tested empirically in an integrated dataset.”

>>In the analysis presented in the new section in the Extended Materials and Methods “3.4 Rate of protein evolution (ω) as a function of effective population size (N_e)” (lines 620 onward), we explicitly integrate the estimates of the two population sizes from the nucleotide differences, assuming the same mutation rate.

>>The results of this new analysis are presented (and interpreted with caution given only two population sizes are used) in the main text (lines 132-141): “As theoretically predicted (Latrille & Lartillot 2021), the rate of purifying selection appears to linearly decrease with the logarithm of the expression rate (Fig. 2A). After also estimating the change in rate of purifying selection (π_N/π_S) as a function of the effective population size of the two penguin species in log scale (Supp. Fig. 8), we show that all estimated slopes are statistically different from zero and negative. However, the slope estimates are not significantly different from each other and their confidence intervals overlap (Supp. Fig. 8). Compatible with the assumptions that proteins are selected for their conformational stability or for protein–protein interaction, these results suggest that the effect of effective population size and gene expression rate can be considered together in integrated models of evolution. However, they should be assessed more thoroughly, by comparing more population sizes.”

**One of the reviewers also noted in reference to the simulation study of population size versus selection that it is difficult to evaluate the "larger effect" of gene expression versus population size because "the two variables are compared on different scales". I agree. In any case, my opinion is that the comparisons of population size versus gene expression could be omitted entirely and the paper would be improved. The other results stand on their own and support the authors' arguments for considering gene expression levels when evaluating selection in populations of interest to conservation biologists, etc.

>> We kindly disagree on this point. When considering the E - R anticorrelation, the toxic accumulation of misfolded proteins has been suggested as a possible explanation. If this is the case, gene expression and effective population size can actually be compared on the “same scale” as has been demonstrated in Latrille and Lartillot 2021. In fact, it is possible to analytically derive the change in selection (ω) as a function of Ne and gene expression (y) as per Equation 18 (in Latrille and Lartillot 2021). From this equation, ω linearly decreases with Ne (in log scale), as well as with y (in log scale). Importantly the slope of the linear model is the same for both.

>>Excitingly, one of the authors of Latrille and Lartillot (2021) contacted us after reading our manuscript in biorxiv asking whether it would be possible to test their theoretical model with our empirical data. We were glad to agree, believing the manuscript would be largely improved and the new analyses would also address some of the comments from the Recommender and the Reviewers.

>>As a result, the manuscript now includes the new sections 3 “Testing the effect of gene expression and population size on purifying selection”, where we present the theoretical model and the new analyses included in the revised version of the manuscript (lines 539 onward).

>>This additional analysis has now been integrated into the main text (lines 50-55 and 132-141), as detailed in the answer to the previous comment.

**Please respond to all the reviewers' comments if you choose to revise your paper for reconsideration.

>>We carefully addressed all of the reviewers’ comments. Please find our point-by-point reply below.

Review by Tanja Pyhäjärvi

**Trucchi et al combine genetic polymorphism data and gene expression data of two penguin species (King and Emperor) to examine the effects of gene expression level and effective population size (Ne) on the level of purifying selection.

**The manuscript seeks to demonstrate the relationship between gene expression level and purifying selection in two species with different Ne. However, the method used to infer the effect of purifying selection, the ratio of synonymous vs. nonsynonymous segregating sites, is the weak link of the work. The data could be used to actually estimate piN/piS, a more widely used and less biased measure of the extent of purifying selection. Since this is a very essential estimate for the conclusions, it would be important to obtain as unbiased measure as possible.

>>We thank the reviewer for pointing out this potential issue with our estimates. We now use estimates of piN/piS per gene, as detailed in the new section 3.2 in the Extended Materials and Methods, lines 558 onward), to demonstrate the anticorrelation between gene expression level and purifying selection. Importantly, the new results are fully in line with the results obtained previously by using the ratio of synonymous vs. nonsynonymous segregating sites, suggesting the latter are robust enough and can be employed in other studies if piN/piS are not available.

**In addition, to state that gene expression has larger effect than effective population size is an overstatement, given that only two very closely related species have been studied here. Wouldn't a more fair comparison be to compare the effect of gene expression level across all genes to the effect of Ne variation across all possible Ne:s? It is also very essential in the text to clearly separate the distribution of s, or its shape from the distribution of Ne*s. The gene expression level and its distribution could act as a proxy to the former, but not the latter as it, by definition, ignores the differences in Ne.

>>See our reply to the Recommender's comment above about comparing only two population sizes.

>>We fully agree with the reviewer that gene expression level is a proxy of s and not of Ns. We have now made this much clearer in the text (lines 46-47, 63-66)

**In several places it is stated that evolutionary rate and gene expression anticorrelation has not been estimated in natural populations. It would be fair to cite and summarize findings of e.g., Slotte et al. (2009, global sample of Arabidopsis accessions, https://doi.org/10.1093/gbe/evr094), Josephs et al. (2017, Capsella grandiflora sample from a natural population https://doi.org/10.1093/gbe/evx068) or Galtier et al. 2016 (2016, 44 non-model animal species https://doi.org/10.1371/journal.pgen.1005774) just to name some that have observed the relationship. If the authors were referring to gene expression, not the polymorphism data from natural population, this should be clarified as they inform about very different phenomena (e.g., protein misfolding in laboratory vs. natural selection in laboratory populations).

>>The reviewer is right as we missed references to these relevant previous studies. We amended by rephrasing the text to account for their results.

>>Of note, while Slotte et al (2009) and Joseph et al (2017) are still focused on E - R anticorrelation in terms of protein evolution (dN/dS) and not standing polymorphism (pN/pS), Galtier et al (2016) found a significant difference in across-species average πN/πS between highly and lowly expressed genes using 44 different species. The latter was added to a list of

studies, suggested by the other reviewers, focusing also on population level diversity. See also our reply below.

>>We rephrased the text at lines 56-65 as follows:
“Evidence for E-R anticorrelation has been found in several interspecific comparisons by estimating fixation rates (*d*) of nonsynonymous (*N*) over synonymous (*S*) mutations (i.e., *dN/dS)* in genes with different expression rates (Slotte et al 2011, Zhang and Yang 2015, Joseph et al 2017). Considering diversity at the population level, E-R anticorrelation should explain differences in nonsynonymous and synonymous segregating polymorphisms (*p*) across genes (i.e., *pN/pS* or as the corrected estimate *πN/πS)*. Although such a pattern has been observed in a few wild populations (Carneiro et al 2012, Williamson et al 2014, Hodgins et al 2016, Galtier et al 2016), recent laboratory experiments on model organisms have instead provided contrasting results (Wu et al 2022, Shibai et al 2022). More importantly, the relative contribution of gene expression and effective population size to purifying selection has not been empirically explored.”

>>We have also changed the text in the abstract as follows:
“However, estimates of the effect of gene expression on segregating deleterious variants in natural populations are scarce” (line 23-24).

**This may be unnecessary for pre-prints, but I would personally prefer the methods to be part of the main manuscript text, not as a separate section. Also, the extent of materials and methods seems very lengthy in comparison to the other parts of the preprint. Several key concepts, hypotheses and conclusions are in the extended methods section. The preprint could be improved by bringing some of that content to the main text and making the style of writing and presentation more coherent.

>>Following the Recommender’s advice, we prefer to keep the Methods as an Extended Methods section following the Main text (see our Reply above).

>>However, following one of the detailed comments below we moved to the Main text, at the end of the Introduction, the hypotheses which were formerly stated in the Extended Methods section (lines 73-85; see below for details).

Detailed comments:

**Line 50: And/or because highly and widely expressed genes have conserved essential functions?

>> We have now included a reference to a recent review (Bédard et al 2022) where the different hypotheses about the cause of E - R anticorrelation have been discussed in the light of available empirical evidence (line 50).

>>Bédard, C., Cisneros, A. F., Jordan, D., & Landry, C. R. (2022). Correlation between protein abundance and sequence conservation: what do recent experiments say?. *Current Opinion in Genetics & Development*, *77*, 101984.

**L. 52-57 this statement is not completely fair description of the current literature.

>>This comment refers to one of the main comments above. We changed the text at lines 56-65 to better account for the current literature. See our detailed reply above.

**L. 61: Of course not, because the effect depends on the joint effect of Ne and s. If the idea is that the relationship is something that is not linear, please clarify.

>>The reviewer is right. Our questions were not properly formulated. Our question here is: at a certain small population size, will the product of Ne x s be <1 for genes with low expression rate while it stays much >1 for genes characterised by high expression? In other words, highly expressed genes could be subjected to considerably higher selection coefficients so that they will not be prone to accumulating deleterious mutations down to very small population sizes.

>>We re-phrased these questions (lines 65-70) as:
“Theory predicts that the efficiency of purifying selection depends on the product of effective population size and selection coefficient to be much larger than 1. We can therefore ask whether genes with high expression levels are characterised by large enough selection coefficients so that purifying selection still exerts its effect even when populations are small. On the other hand, understanding the range of selection coefficient values across genes would help identify those genes which are more vulnerable to decreasing population size.”

**L. 67-68: This statement needs to be more specific. Effect in what sense? At what range of Ne variation?

>>We agree with the reviewer. This statement was thought as the summary of this result in the previous version of the manuscript “Average pN/pS across genes binned in 5% percentiles of expression rate shows a declining trend of pN/pS with expression, with the bin average dropping by *ca.* 80% across the whole range of gene expression in both species (average pN/pS: from > 1.1 in the bottom 5%, to < 0.4 in the top 5% of expression rate; Fig. 2A). Conversely, the difference in pN/pS between the two species (*i.e.*, likely due to the effect of the population size) spans from 2% to 37% across the whole range of gene expression, suggesting that effective population size is a less important determinant of pN/pS in these species.”

>>However, following another suggestion from the same Reviewer we replaced the text formerly at lines 67-68 with the description of the main questions and expectations tested in our study (lines 73-85).

**Figure 1 B. Please check the color scheme. It is not clear how to separate E vs. K from non-synonymous to synonymous polymorphism and divergence.

>>Agreed. The Emperor penguin divergence data is shown in teal (filled shape), while the King penguin in orange (filled shape). In the case of panel 1B, non-synonymous are now shown with an empty shape, while synonymous are shown as a lighter shade of the “species” colours as for panel 1A.

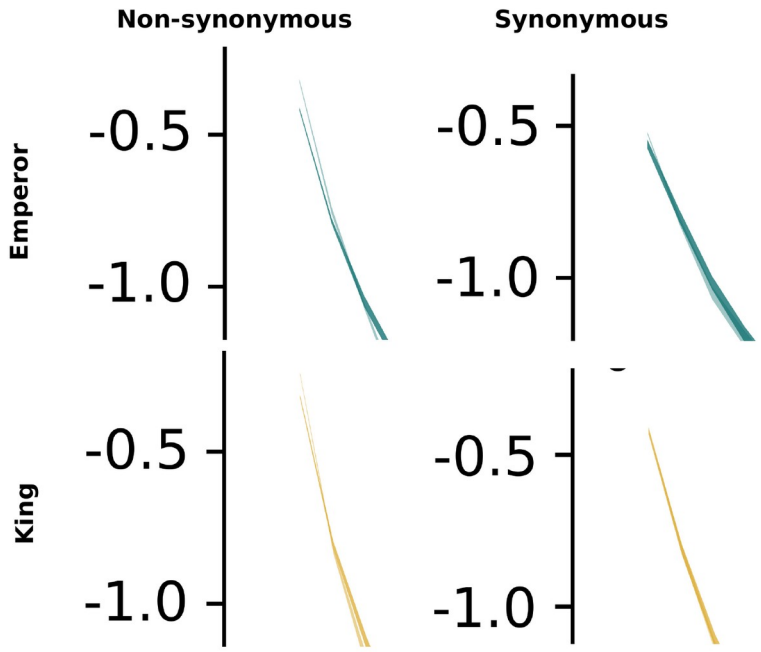
**Figure 2C: Counts of segregating sites are prone to bias in mappable genome. Pi estimates would be better for comparative purpose as they take account differences in the amount of monomorphic synonymous and nonsynonymous sites as in the sample size.

>>Done and all results hold. See our reply to one of the main comments above and to other comments below.

**Figure 3 would be much easier to evaluate if it would not be log10 transformed and if it would be presented as a histogram of allele frequencies in all frequency classes (allele frequency spectrum). In the current version, it remains puzzling e.g., how nonsynonymous spectrum > CPM 0.3 is consistently smaller than the synonymous, across all frequency classes. When obtaining the AFS from NGS data, it is critical to explain how missing data was handled as it results in variation in maximum derived allele count among sites. This aspect is critical for the main conclusions of the paper and thus needs to be explained in detail.

>>If the y-axis of AFS was not log10 transformed, the vast majority of it would be taken by the allele counts of 1, whose proportion largely overwhelms the other count classes. Such a log-scale has been commonly applied in previous studies for visualisation of the AFS.

>>As each allele count class in the AFS is given as its relative frequency to the total, nonsynonymous spectrum > CPM 0.3 cannot be consistently smaller than the synonymous. In fact, the non-synonymous AFS is higher than the synonymous one for the small allele counts (see below). We agree with the reviewer that this pattern is not visible in Figure 3 however we already put a note about this issue in the legend of the figure (lines 168-169).



>>As we were aware that missing data are difficult to handle when building AFS, we only selected loci without missing data as explained in Extended Methods Section 1.5 (now called “Final data filtering and sanity checks”; lines 429-430). Given the high coverage of our dataset, such strict filtering did not impact too much the number of loci used.

**Figure 4, Are the ratios not calculated per synonymous sites or nonsynonymous sites? Are these ratios only based on counts of segregating sites? I strongly recommend using piN/piS instead of just counts of segregating sites, where there are more clear expectations and earlier empirical evidence to compare your results to.

>>In Figure 4 we show the results of the simulations described in the Extended Methods Section 4 (now called “Estimating the selection coefficients of highly expressed genes using realistic forward simulations”). In the case of the simulations, the length of the coding sequence is fixed as well as the ratio of occurrence of deleterious (non-synonymous) and neutral (synonymous) mutations (2.31:1 as suggested in Kim et al 2017) but there is no “real” open reading frame. The ratio above (2.31 : 1) is used to scale the opportunities of N or S. As there is no missing data either, estimates of pN/pS from simulated data are expected to be unbiased. Following a previous comment, we re-analysed the empirical data using piN/piS showing that the resulting patterns were the same as when using pN/pS.

**L. 326 Minimum depth of 3 reads per individual seems very low. It is quite easy to miscall heterozygotes and homozygotes with three reads. I suggest using much higher depth threshold at genotype level.

>>This is only the first level of filtering in order to produce a “starting-point” dataset to be publicly released. Loci are then further filtered according to downstream analyses. In all our analyses, we explain in Extended Methods Section 1.5, lines 430-432, that we filtered loci to have an average coverage per allele across samples between 6X and 8X (one SD more or less than the mean allele coverage, 7X; between 12X and 16X per genotype).

**L. 332 vcf files should be available in a repository.

>>We agree but VCF files are not usually uploaded to the NCBI or ENA database (raw sequencing reads are available there) and they are too large for Dryad or Zenodo repositories. The final clean dataset (daf.joint.no00; lines 427-428) was made publicly available in Zenodo (10.5281/zenodo.10688854) but all vcf files will be made available upon request.

**L. 342 Ancestral is not equal to reference and derived not equal to alt allele. The whole section 1.4 should be written as a scientific text and not as a list that is somewhat hard to interpret. Clearly explain here what was done. This is an essential part of the analysis and needs to be clearer.

>>After assessing the ancestral and derived alleles on the basis of the algorithm explained in the Supplementary Figure 1 and implemented in the script vcf2missenseFreq.2d.py (available at https://github.com/emitruc/ExpressionLoad), we labelled the ancestral allele as *ref* and the derived allele as *alt* in the daf.joint dataset.

>>The daf.joint dataset is a custom-made table and the text at lines 403-415 describes each column field. So, it has to be considered as an explanation of the dataset table only. We added the following text at line 400: “column labels are in brackets”

**L. 336 Why were the King and Emperor allele counts summed up?

>>This was a simple typo. Of course, the King and Emperor allele counts were not summed up. We change the text as follows (line 399): “We calculated the joint derived allele counts for King and Emperor…”.

**Section 1.5 contains important details of polarizing the SNPs in interpreting the data. Part of the text belongs to the main text results and discussion. As a whole, this section would benefit from an introductory paragraph explaining why this procedure is necessary and it could be combined with section 1.4 There are vague references to population genetic theory, but the exact predictions should be stated, and relevant literature cited. Sex chromosomes and HWE are passingly mentioned but not really put into context.

>>We agree with the Reviewer that the title of this section was misleading. However, Section 1.5 does not include any details about SNP polarisation. In fact, the algorithm used for inferring the derived allele was presented in Section 1.4 (We now add more details in that section - lines 395-400) as well as the resulting dataset (Supp Table 1). In Section 1.5 we describe further filters applied to this dataset: SNPs which are monomorphic in both the Emperor and King are excluded (different only in the outgroup species); SNPs are also excluded if difficult to polarise (flagPol filter), and if showing any missing data (no missing data was allowed); finally, we selected loci with an average sequencing coverage range between 6X and 8X per allele (12X - 16X per genotype) across all samples.

>>We then provide an empirical assessment for our very strict coverage range threshold in Supp. Fig. 2. As we already removed sex-related scaffolds from our dataset, we believe that some sex-chromosome related regions have been incorrectly assembled into autosomal

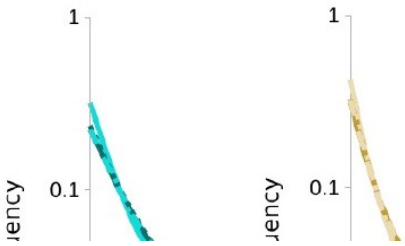
scaffolds as the sex-chromosome related bias in Supp. Fig. 2 corresponds to the male:female ratio in the samples. However, these regions also show biases in the average allele coverage due to the non 50:50 presence in the sample. Moreover, we detect an excess of loci with 50% heterozygosity in both species (Duplicated regions in Supp. Fig 2); at closer inspections, those SNPs resulted as heterozygous in all individuals and with higher coverage than genomic average. As these could be incorrectly resolved duplicated regions in the reference genome assembly, we minimised their contribution to our dataset by applying a stringent coverage filter.

>>The final part of Section 1.5 is a further sanity check of the dataset showing the site frequency spectra for different classes of loci (intergenic, intronic, synonymous and missense) after applying the filters mentioned above. All patterns in Supp. Fig. 3 are in line with the expectations from classic population genetics theory and genomic structure.

>>Therefore, describing filtering strategies and data sanity checks, we believe Section 1.5 can be deemed as part of the Materials and Methods. We understand that the title could have been misleading, hence we re-phrased as: “1.5 Final data filtering and sanity checks” (line 425).

**Supplementary figure 3: see comments on Figure 3

>>See our answer above. As each allele count class in the AFS is given as its relative frequency to the total, none of the spectra could be consistently smaller than another one. In fact, the missense AFS are higher than the other spectra for the small allele counts. In this case, the difference is visible in supplementary figure 3 (solid light blue or light orange line for Emperor and King penguin, respectively).



**1.6. How does the vcftools handle missing data when estimating pi? It may assume that all missing sites are invariant. See for example Korunes and Samuk 2021 for possible pi estimate biases of vcftools (https://doi.org/10.1111/1755-0998.13326).

>>Yes, the reviewer is right, *vcftools* considers missing data as invariant. For low coverage data or when structural variation is high at the intrapopulation level (i.e. when missingness across samples is high) a better alternative for handling missing data when estimating molecular summary statistics in genomic windows is, of course, *pixy.* We were aware of this issue since the beginning of this project and tested *pixy* alongside *vcftools* recording negligible differences likely due to the medium-high coverage of our dataset.

**L.466-471 and throughout the manuscript: GitHub or other repository may be a better place to share the exact code that was used to produce the data.

>>All of the custom scripts used in our analyses were already available in github (github.com/emitruc/ExpressionLoad, github.com/ThibaultLatrille/PenguinExpression, and github.com/PiergiorgioMassa/penguin_gene_expression_slimulations). At these lines, there

are some simple bash terminal commands which are provided for sake of clarity. We think there is no need to put them in github.

**L477-478 Please justify why counts of synonymous and missense polymorphic sites are used, rather than pi (the mean pairwise differences per bp). Further, more appropriate than normalizing by the CDS length would be to calculate the amount of total synonymous sites (or 4-fold sites, which is more straightforward) and use that as a nominator to obtain per bp estimates of nucleotide diversity in different SNP categories. Further, these estimates must be adjusted according to the same or similar filtering criteria that were used for SNPs. Just normalizing by CDS length does not consider that not all nucleotides of the CDS are part of your data and that unequal proportion of them are synonymous and missense.

>>We now include the results using piN and piS for which normalisation is performed according to the total opportunities for synonymous and non-synonymous mutations (Section 3.2 in the Extended Methods). At any rate, the data filtering before piN and piS estimated is the same as before (daf.joint dataset; see above). See our reply above for further details.

**L486: these hypotheses would better fit to the main text of the preprint.

>>We agree with the Reviewer and moved the hypotheses from the Extended Materials and Methods section to the main text and integrated them at the end of the Introduction with a few modifications (lines 73-85):

“First, if the selection coefficient of a gene is mainly determined by its expression rate, we should observe a decline in the effect of purifying selection (e.g. πN/πS) with increasing expression rate and such decline should be determined by a corresponding decline in missense polymorphism only. Our second question concerns the relative weight of population size (Ne) and gene expression (s) in driving purifying selection. When comparing populations of different sizes, smaller populations show lower diversity at both neutral and deleterious sites, but higher πN/πS because of larger drift which reduces the efficacy of purifying selection. If population size is the main driver of purifying selection (1/Ne > s across the whole range of gene expression), we expect that both the diversity and the πN/πS differences between the two populations of different sizes will be the same across the whole range of gene expression. Conversely, if high gene expression is the main driver of purifying selection (s > 1/Ne for highly expressed genes), we expect the difference in diversity between the two populations of different size to decline with increasing gene expression rate for deleterious sites but not for neutral ones.”

**L. 495-497 “difference will be the same” need more explanation. Difference measured how? Please tie this to the prediction that the effect of selection depends on the product of Ne and s. Please provide the equations to clarify the prediction. Do you suggest that Ne does not have an effect at all?

>>See our reply to the previous comment.

****Review by anonymous reviewer 1, 06 Sep 2023 20:50**
(Note that this review was jointly performed by two people)

**This manuscript investigates the correlation between gene expression and measures of purifying selection, primarily pN/pS, in two separate penguin populations, along with investigating the effect of increases in purifying selection vs increases in population size on pN/pS. These are both interesting questions to investigate and have clear importance for questions regarding protein evolution. The use of wild transcriptome data to investigate the polymorphism vs expression relationship is notable. The main claim of the study is that gene expression is a stronger driver of purifying selection than population size in this system. The manuscript also argues that gene expression levels can approximate the distribution of fitness effects in non-model species. We found that this work is overall interesting, but have a few concerns about the statistical analyses, population genetics mechanisms, and claims about the novelty of the study, that we discuss below.

>>We thank the reviewers for the positive considerations of our work and we addressed all of their comments to further strengthen its soundness and implications.

Major comments:

**1. We are concerned about the choice to use binned data to estimate the difference of nonsynonymous and synonymous polymorphisms across expression levels (Fig 2 and the results section titled “Purifying selection more efficiently removes nonsynonymous segregating variants in genes while expression rate increases”). Since these two variables are naturally continuous, it is more appropriate to analyze them as scatterplots instead of arbitrarily binning them, potentially inflating the statistical signal. We suggest re-plotting figure 2 as a scatterplot. There may be outliers along the expression dimension, which could be why the authors binned their expression values into percentiles, but they could also look at the logarithm of expression to alleviate this problem while keeping the variable continuous. The authors would then calculate a spearman’s correlation between pN/pS and log(gene expression + 1)

>>We completely understand this concern. Binning genes by expression value and presenting the bins as boxplot was chosen to better visualise the anti correlation pattern together with the variance in each bin. However, we agree that the two variables are naturally continuous and that it could be more appropriate to visualise them without binning (as a scatterplot). We now present the results without any binning, as well as with a decreasing number of bins from 100, 50, and 20, and compute the slope (χ) and the fit (R^2) of the linear regression (Extended Materials and Methods section 3, Supp. Fig. 6). Of note, we now also use $\pi N/\pi S$ instead of pN/pS to estimate purifying selection on segregating polymorphism and we also add the analyses using dN/dS to estimate purifying selection on species genetic distance. To compare our results to Zhang and Yang (2015), we also plot both the selection effect (as $\pi N/\pi S$ or dN/dS) and the gene expression in log scale (Supp. Fig. 7). However, we still believe that the plot with genes binned by 5% percentiles of expression rate provides a better visualisation of the E -R pattern (Fig. 2A in the main text), but we are open to discuss this further.

>>The slope of $\pi N/\pi S$ as a function of log expression level is not dependent on the number of bins used to compute $\pi N/\pi S$ or dN/dS. However, for fewer bins, the linear model is a strong fit (high R2), but, of course, the fit decreases as the number of bins increases.

**2. The authors show in Figure 1 that they have dN/dS measurements for each species, but they only focus on pN/pS. We were curious whether the dN/dS results recapitulate the same trends as pN/pS, seeing as how the two species don’t seem to differ drastically in dN/dS. Some additional explanation on why only pN/pS results are presented would be appreciated,

since dN/dS also quantifies purifying selection. In addition, having dN/dS results displayed more prominently would make this study easier to compare to the many previous studies that have looked at the relationship between expression and dN/dS.

>>We did not show dN/dS in the first version of the manuscript for two reasons: first, being the divergence between the two species quite shallow, there are not many fixed differences per gene so that dN/dS estimates have more uncertainty. Second, the anti correlation between gene expression and purifying selection has been widely acknowledged at the interspecific level (dN/dS) but it is not yet clear how strong it could be at the population level, that is on patterns of segregating polymorphism pN/pS (or πN/πS). So, we believe the latter test encapsulates the novelty of our study. At any rate, we now add the analyses using dN/dS and show that the anti correlation pattern is also present at the interspecific level (Extended Materials and Methods section 3, Supp. Fig. 6, 7).

**3. One of the study's main claims is that gene expression has a larger effect on purifying selection than changes in population size. However, it is hard to evaluate this claim because these two variables are compared on different scales with different units and different scopes. For example, is a change in height by 5 inches comparable to a change in weight by 5 pounds? Similarly, is a decrease in selection coefficient from -0.1 to -0.01 comparable to a population size change from 100,000 to 10,000? To compare the effects of the two different variables, it would be helpful to standardize them according to their respective mean and variance. We realize this might not be possible for the natural data, but it could be helpful for the simulated data. Alternatively, it could be helpful to look at population scaled selection coefficients (2*Ne*s for diploids) instead to demonstrate this claim more clearly.

>>We kindly disagree with the reviewers here. In the context of the E - R anticorrelation, gene expression and effective population size can actually be compared on the same log scale as it has been demonstrated in Latrille and Lartillot 2021. In fact, it is possible to derive analytically the change in selection (ω) as a function of Ne and gene expression level (y) as shown with equation 18 (Latrille and Lartillot 2021, eq.18). From this equation, ω is linearly decreasing with Ne (in log scale) as well as with y (in log scale), importantly the slope of the linear model is the same for both. We completely revised this part of the analyses after engaging in a collaboration with the authors of the study mentioned above who contacted us to use our empirical dataset to test their predictions. See the new section 3 in the Extended Materials and Methods.

**4. While it is clear that gene expression is highly correlated with measures of purifying selection, and thus could be used as a proxy for purifying selection, we are not sure if gene expression could approximate the entire distribution of fitness effects based on the data presented here. A DFE includes information about both the mean and variance of mutation effects. We can see how gene expression could provide information about the mean of the DFE (higher average expression, lower average selection coefficient), but we are not clear how it provides information about the variance. Unless perhaps the mean and variance are correlated or linked somehow? We would appreciate either some clarification on this point or rewording of the claim.

>>That is indeed a good point raised by the reviewers. We can approximate the average DFE with the level of gene expression but we do not have much information on DFE variance. As we were aware of this issue we preferred to use the concept of "gene expression coefficients" instead of DFE. For the sake of clarity, we remove the brackets "(i.e., distribution of fitness effects)" that was in the abstract. DFE is not used elsewhere in the text.

**5. The authors collected gene expression data across multiple tissues, so we assume that the gene expression levels in their plots show expression averaged across all sampled

tissues. We couldn't find this detail stated explicitly though, so we would appreciate some clarification on this. In addition, we don't want to require additional analyses but wanted to suggest for here or future work investigating how tissue-specificity of expression also relates to purifying selection, since the authors may have that data already? Tissue-specificity is typically highly correlated with average expression levels (For example, see Slotte et al 2011: https://doi.org/10.1093/gbe/evr094) and Duret and Monod 2000 is cited in the introduction which was one of the earlier papers to demonstrate the importance of tissue-specific expression on evolutionary rates.

>>Yes, correct: we used global gene expression data across five tissues. It is mentioned in the Method section (lines 506-509: "As the target of our study was to estimate the global level of gene expression (across tissues), a total of six RNA pools (three pools of five tissues per three individuals for each species) were assembled starting from 15 RNA samples per species, after concentration was normalised."). We also have tissue-specific RNAseq data, although from a different sequencing approach (5 tissues x 10 individuals x 2 species, 3'-end sequencing; just released as a pre-print (doi.org/10.1101/2023.11.29.569211) and under review in Molecular Ecology journal). The reason to use global expression instead of tissue specific was motivated by the Omnigenic model (Boyle et al 2017) which claims that important genes are those with globally high expression level across multiple tissues. Of course, we performed some preliminary analyses on the data with different tissue samples and checked that the anticorrelation pattern is still there. As the present study is already quite dense, we prefer not to add more analyses but we are open to further discuss this issue with the reviewers in case they think the tissue-specific dataset would highly improve the work.

**6. This study includes two different penguin species, Aptenodytes patagonicus and Aptenodytes forsteri, and genotypes were identified by aligning reads in both species to the same reference genome (Aptenodytes forsteri) (Extended methods section 1.3). Presumably, reads from A. forsteri will align at a higher rate and lead to more genotype calls compared to A. patagonicus. Is it possible that this reference bias could explain some of the results of this study?

>>Given the very shallow divergence between the two *Aptenodytes* species, we did not expect any relevant bias using the reference genome of one species to map the DNA data of the other species. We estimated the genetic divergence between the two species as less than 1% on a 2kb region at 3' UTR, which is not deemed as a concern for cross-reference mapping.

>>One possible bias could be that lowly expressed genes accumulate more changes that, in turn, would cause lower cross-specific reference RNAseq data mapping and, hence, even lower estimates of gene expression for the non-reference species. However, we showed that both the genetic diversity and the gene expression is highly correlated between the two species (main Fig. 1 and Supp. Fig 5), suggesting a negligible bias by lower mapping rate of *A. patagonicus* data to the *A. forsteri* genome.

>>We are currently assembling a highly-contiguous and much better annotated reference genome for the *A. patagonicus* and we would be up to replicate these analyses as soon as it is finalised.

**7. This manuscript emphasizes that it is the first to investigate selection on genes of different expression levels in natural populations. However, there are many studies that use genotypes from natural populations with expression from lab-reared individuals to address the relationship between gene expression and selection. For example see.
Carneiro et al. 2012: https://doi.org/10.1093/molbev/mss025
Williamson et al 2014 https://doi.org/10.1371/journal.pgen.1004622

Hodgins et al. 2016 https://doi.org/10.1093/molbev/msw032
If the authors mean to imply that the novelty of this study comes from using wild-collected transcriptome data, it would be useful to know how their transcriptome data compares (and differs) from expression data from captive or lab-reared individuals or about their expectations for why transcriptomes from wild-caught individuals will differ from those of lab-reared individuals.

>>The reviewers are right as we missed references to some relevant previous studies. We amended by rephrasing the text to account for their results.

>>Although there could be some biases in using captive or lab-reared individuals to gather global expression data, we also think that the overall signature of E-R anticorrelation should hold. On the other hand, such studies were still mainly focused on genetic divergence data (dN/dS), whereas estimates at the population level (using site frequency spectra or piN/piS) appears as less refined. For example, in some of these studies, the anticorrelation was shown by coarsely grouping genes in four expression categories.

>>We rephrased the text at lines 56-65 as follows:
“Evidence for E-R anticorrelation has been found in several interspecific comparisons by estimating fixation rates (*d*) of nonsynonymous (*N*) over synonymous (*S*) mutations (i.e., *dN/dS)* in genes with different expression rates (Slotte et al 2011, Zhang and Yang 2015, Joseph et al 2017). Considering diversity at the population level, E-R anticorrelation should explain differences in nonsynonymous and synonymous segregating polymorphisms (*p*) across genes (i.e., *pN/pS* or as the corrected estimate *πN/πS)*. Although such a pattern has been observed in a few wild populations (Carneiro et al 2012, Williamson et al 2014, Hodgins et al 2016, Galtier et al 2016), recent laboratory experiments on model organisms have instead provided contrasting results (Wu et al 2022, Shibai et al 2022). More importantly, the relative contribution of gene expression and effective population size to purifying selection has not been empirically explored.”

>>We have also changed the text in the abstract as follows:
“However, estimates of the effect of gene expression on segregating deleterious variants in natural populations are scarce” (line 23-24).

Minor comments:

**Supplemental section 1.3: Annotated variant files are said to be available upon request. It would be nice if these were deposited somewhere once the manuscript is accepted for publication.

>>We agree but VCF files are not usually uploaded to the NCBI or ENA database (raw sequencing reads are available there) and they are too large for Dryad or Zenodo repositories. The final clean dataset (daf.joint.no00; lines 427-428) was made publicly available in Zenodo (10.5281/zenodo.10688854) but all vcf files will be made available upon request.

**Supplementary methods section 5: The definition of genetic load here includes the phrase “cost paid”. We think it would help the reader to break down this phrase a little more and mention the accumulation of deleterious mutations that decrease the fitness of “high load” individuals relative to individuals with fewer such mutations.

>>Agreed. We added the following sentence (lines 733-735):
“Deleterious mutations can accumulate in some individuals as a consequence of small population size (i.e., high genetic drift and high inbreeding) reducing the fitness of these high load individuals as compared to individuals which bear fewer such mutations.”