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.

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)

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.

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.

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.

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.

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?

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..

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.

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.