Molecular Biology and Evolution, 2025, **42**, 1–9
https://doi.org/10.1093/molbev/msaf146
Advance access publication 10 June 2025
Discoveries

MBE OXFORD

High Gene Expression Predicts Extremely Low Segregation of Deleterious Mutations in Large Penguin Populations

Emiliano Trucchi [1,*], Piergiorgio Massa [2], Francesco Giannelli [1], Thibault Latrille [3], Marco Gargano [1], Flavia A. Nitta Fernandes [1,4], Lorena Ancona [1], Nils Chr Stenseth [5], Joan Ferrer Obiol [6,7], Josephine Paris [1], Giorgio Bertorelle [8], Céline Le Bohec [4,9,10]

[1]Department of Life and Environmental Sciences, Marche Polytechnic University, Ancona, Italy
[2]Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum University of Bologna, Ravenna, Italy
[3]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland
[4]Université de Strasbourg, CNRS, IPHC UMR 7178, F-67000, Strasbourg, France
[5]Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway
[6]Department of Environmental Science and Policy, University of Milan, Milan, Italy
[7]Department of Biology, Colorado State University, Fort Collins, CO, USA
[8]Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara, Italy
[9]CEFE, Université de Montpellier, CNRS, EPHE, IRD, Montpellier, France
[10]Département de Biologie Polaire, Centre Scientifique de Monaco, Monaco, Principality of Monaco

***Corresponding author:** E-mail: e.trucchi@univpm.it.
Associate editor: Xuming Zhou

Abstract

Purifying selection is the most pervasive type of selection, as it constantly removes deleterious mutations arising in populations, directly scaling with population size. Highly expressed genes appear to accumulate fewer nonsynonymous mutations between divergent species' lineages (known as E–R anticorrelation), pointing toward gene expression as an additional component modulating the selection coefficient of protein-coding mutations. However, estimates of the effect of gene expression on segregating deleterious variants in natural populations are scarce, as is an understanding of the relative contribution of population size and gene expression to purifying selection. Here, we analyze genomic and transcriptomic data from two natural populations of closely related sister species with different demographic histories, the Emperor penguin (*Aptenodytes forsteri*) and the King penguin (*Aptenodytes patagonicus*), and show that purifying selection at the population level depends on gene expression rate, resulting in very high selection coefficients at highly expressed genes. Leveraging realistic forward simulations, we estimate that the top 10% of the most highly expressed genes in a genome experience a selection pressure corresponding to an average selection coefficient of −0.1, which decreases to a selection coefficient of −0.01 for the top 50%. Gene expression rate can be regarded as a fundamental parameter of protein evolution in natural populations, maintaining selection effective even at small population size. We suggest gene expression could be considered as a major component of gene-specific selection coefficients, which are notoriously difficult to derive in nonmodel species under real-world conditions.

Keywords: purifying selection, gene expression, population size, protein evolution, forward simulations

Introduction

Protein evolution is constrained by purifying selection, which prevents changes in the underlying gene sequence with a deleterious effect on organismal fitness from spreading in natural populations. The intensity of purifying selection on deleterious mutations is directly correlated with the effective size of a population (N_e; Charlesworth 2009; Akashi et al. 2012), determined by species-specific life history traits and population-specific demographic trajectories (Figuet et al. 2016 ; Chen et al. 2017), and with the selection coefficient (s) of each mutation. However, genes with a globally high expression rate across tissues show a slow rate of accumulation of deleterious substitutions (Duret and Mouchiroud 2000; Pál et al. 2001 ; Zhang and Yang 2015), suggesting high selection coefficients on any mutation appearing therein. Such an inverse correlation between the rate of evolution and gene expression (so-called *E–R anticorrelation*) could be caused by the strong selection acting against the toxic accumulation of misfolded or misinteracting proteins in cells (Yang et al. 2012; Park et al. 2013; Wu et al. 2022, but see Pritykin et al. 2015 and Bédard et al. 2022 for more hypotheses about the causes of E–R anticorrelation). Assuming that proteins are also 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), for each novel mutation altering the protein thermodynamic stability, we can foresee a component of the coefficient of selection which depends on the amount of protein product or, simplifying, on gene expression. The probability of fixation of a new mutation can then be theoretically derived as a function of both gene expression and effective population size (Latrille and Lartillot 2021), but so far the predictions of these models have not been tested empirically in an integrated dataset.

Received: July 3, 2024. **Revised:** April 18, 2025. **Accepted:** May 14, 2025

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. d_N/d_S) 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. p_N/p_S or as the corrected estimate π_N/π_S). Although consistent patterns have been observed in a few wild populations (Carneiro et al. 2012; Williamson et al. 2014; Hodgins et al. 2016; Galtier 2016) and direct competition experiments in *Saccharomyces cerevisiae* show the dependence of the distribution of fitness effects on gene expression level (Wu et al. 2022), other laboratory experiments measuring E–R anticorrelation of de novo mutations in *E. coli* provided not fully conclusive results (Shibai et al. 2022). But note that most laboratory-based evolution experiments have a resolution only for mutations with a selection coefficient ≥ 0.005 (Kofler and Schlötterer 2014). More importantly, the relative contribution of gene expression and effective population size to purifying selection has not been empirically explored. 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 characterized by large enough selection coefficients so that purifying selection still exerts its effect even when populations are small. On the contrary, understanding the range of selection coefficients across genes would help identify those genes which are more sensitive to increasing drift effects in case of decreasing population size.

Here, we use two natural populations of closely related sister species, the Emperor and the King penguins (*Aptenodytes forsteri* and *Aptenodytes patagonicus*), with different demographic histories (Trucchi et al. 2014; Cristofari et al. 2016, 2018), to test the following hypotheses. First, if the selection coefficient of a gene is also dependent on its expression rate, we should observe a decline in the effect of purifying selection (e.g. π_N/π_S) with increasing expression rate and such decline should be determined by a corresponding decline in nonsynonymous polymorphism only. Our second question concerns the relative weight of population size and gene expression in modulating purifying selection. When comparing populations of different sizes, smaller populations show lower diversity at both synonymous and nonsynonymous sites (Fig. 1), but predicted higher π_N/π_S because of larger drift which reduces the efficacy of purifying selection. One possible explanation of the E–R anticorrelation is that, beside their specific function in the cell/organism, proteins can also be selected for their conformational stability. In this model, the selection coefficient of a novel mutation can be decomposed as $s = (s_y\, f(y)) + s_X$, where s_y is the component of the selection coefficient deriving by the changes in the thermodynamic stability of the protein product which can be modified by some monotonic function of gene expression, $f(y)$, and s_X denotes the component of the selection coefficient depending on the actual function of the protein in the cell/organism (of note, y may also have an effect on s_X, which is not included in our simplified formulation). Considering that the population-size-scaled selection coefficient $|N_e\, s|$ has to be $>>1$ for selection to be effective over genetic drift, in the E–R model, the larger the gene expression (y) the higher the contribution of s_y to the overall s. If the contribution of $s_y\, f(y)$ is negligible with respect to the N_e difference between the two penguin species compared here, we predict a linear decline of nonsynonymous diversity in genes with increasing expression level (Fig. 1, top). However, if the contribution of gene expression becomes much larger than N_e, via its contribution to the subterm $s_y\, f(y)$, nonsynonymous diversity should decline to the same level at highly expressed genes regardless of the population size differences (Fig. 1, bottom). Finally, we use realistic forward simulations of evolving populations to estimate the range of selection coefficients producing the same effects of purifying selection as observed in natural populations of Emperor and King penguins.

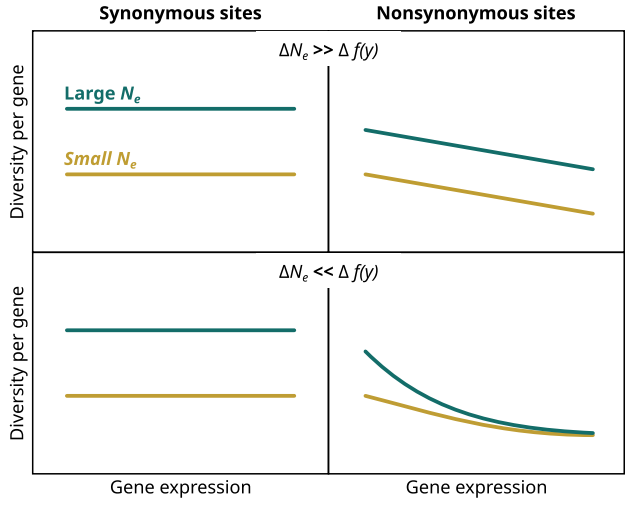

Fig. 1. Sketched diagram of the predicted effects of gene expression on genetic diversity (synonymous and nonsynonymous) under the E–R anticorrelation model. If the difference in population size (N_e) is much larger than the variation in gene expression (y), we expect that the difference in genetic diversity at nonsynonymous sites will remain the same across the whole range of expression rate (top); if variation in y is instead larger than the difference in N_e, we predict the difference in nonsynonymous genetic diversity to decline with increasing gene expression (bottom). Note that the nonlinear decline in the bottom-right plot is purely descriptive of a generic monotonic function of gene expression ($f(y)$).

Results and Discussion

We use high-coverage whole-genome data of 24 individuals per species to estimate patterns of genetic diversity, and whole transcriptome data of five tissues from three young individuals per species to estimate global mRNA expression levels. Young age class was chosen for this study as genes broadly expressed in early life stages have been shown to be the most affected by purifying selection (Cheng and Kirkpatrick 2021). Both Emperor and King penguins feature single, large, and quasi-panmictic populations (Cristofari et al. 2016, 2018), but they show different levels of genetic diversity (Kolmogorov–Smirnov [KS] test P-value $<<0.001$; Fig. 2a), corresponding to their different ecological adaptations and past demographic dynamics after divergence (Cristofari et al. 2016, 2018; Vianna et al. 2020; Cole et al. 2022; Pirri et al. 2022). As a consequence of the historically larger effective population size in the Emperor penguin, this species has a higher proportion of segregating variants, a lower proportion of fixed derived variants, and a lower proportion of segregating nonsynonymous over synonymous variants (KS test P-value $<<0.001$; Fig. 2b); however, the two sister species show a minor difference in the proportion of fixed nonsynonymous over synonymous differences, given their relatively short time since species divergence (KS

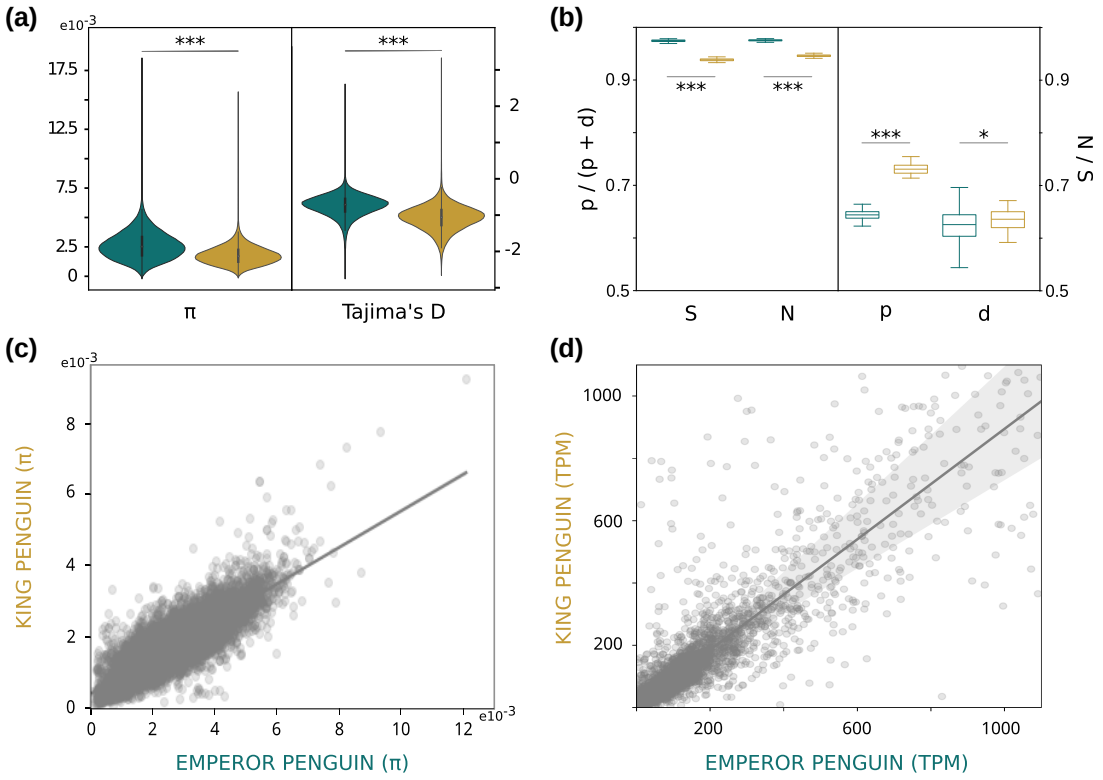

Fig. 2. Patterns of genetic diversity and gene expression in Emperor (teal) and King (gold) penguins. a) Distribution of nucleotide diversity (π) and Tajima's D in 50 kb genomic windows (statistical differentiation assessed via KS test: * < 0.05, ** < 0.01, *** < 0.001); b) block bootstrap (see supplementary methods S3.2, Supplementary Material online) proportion of segregating variants (p) over total segregating and fixed differences (d) at synonymous (S) and nonsynonymous (N) sites (left panel) and estimates of pN/pS and dN/dS (right panel); statistical differentiation assessed via KS test; c) per gene comparison of nucleotide diversity between King and Emperor penguins (Pearson's r: 0.84, P-value < 1E−10); d) per gene comparison of expression rate between King and Emperor penguins (Pearson's r: 0.70, P-value < 1E−10), quantified as TPM (up to TPM = 1,100; see supplementary fig. S5, Supplementary Material online for whole expression range).

test P-value = 0.016). Both gene-by-gene estimates of diversity (nucleotide diversity: π) and expression rate (normalized as transcripts per million [TPM]) are highly correlated between the two species (Fig. 2c and d), thus minimizing any confounding effect of sequence and expression divergence in our downstream analyses.

Purifying selection more efficiently removes nonsynonymous segregating variants in genes while expression rate increases

Corrected estimate of purifying selection on segregating variants per gene, π_N/π_S, clearly declines with increasing gene expression rate (Fig. 3a), dropping by 70% to 80% across the whole range of gene expression in both species. These results hold regardless of binning or not the genes in percentiles of expression rate (supplementary fig. S6, Supplementary Material online) and are consistent with the E–R anticorrelation found in several taxa at the interspecific divergence level as shown in Zhang and Yang (2015) (supplementary fig. S7, Supplementary Material online). As expected, also the rate of fixation of nonsynonymous over synonymous mutations (d_N/d_S) declines with gene expression rate in both species, even if divergence estimates are null for many genes given the shallow split time between two penguin species (supplementary figs. S6 and S7, Supplementary Material online). E–R anticorrelation appears also if we analyze the expression rate of segregating sites across all genes together (mRNA sequencing coverage per site normalized as count per million reads [CPM]) in order to take into account heterogeneous expression rate among exons: again, counts of nonsynonymous over synonymous variants in bins of 0.05 CPM, from 0 to 5 CPM, are inversely correlated with expression rate (supplementary fig. S9, Supplementary Material online). The decline of π_N/π_S with increasing gene expression rate is due to the decreasing count of nonsynonymous variants in highly expressed genes, whereas the count of synonymous variants is stable across the whole gene expression range in both species (Fig. 3b and c). More importantly, the difference in the counts of synonymous variants between the two penguin populations is also stable and always significant (KS test P-value <<0.005), whereas the difference in the counts of nonsynonymous variants decreases with increasing gene expression, with this difference disappearing in the upper 50% to 60% of gene expression rate (Fig. 3c). This result supports the hypothesis that gene expression variation has a larger contribution than population size difference to purifying selection (see Fig. 1, bottom panels). Highly expressed genes are then expected to show very large selection coefficients ($s >> 1/Ne$). As theoretically predicted (Latrille and Lartillot 2021), the rate of purifying selection appears to linearly decrease with the logarithm of the expression rate (Fig. 3a). After 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, we show that all estimated slopes are statistically different from zero and negative (supplementary fig. S8, Supplementary Material online). However, the slope estimates are not significantly different from each other and their CIs overlap (supplementary fig. S8, Supplementary Material online). Compatible with the

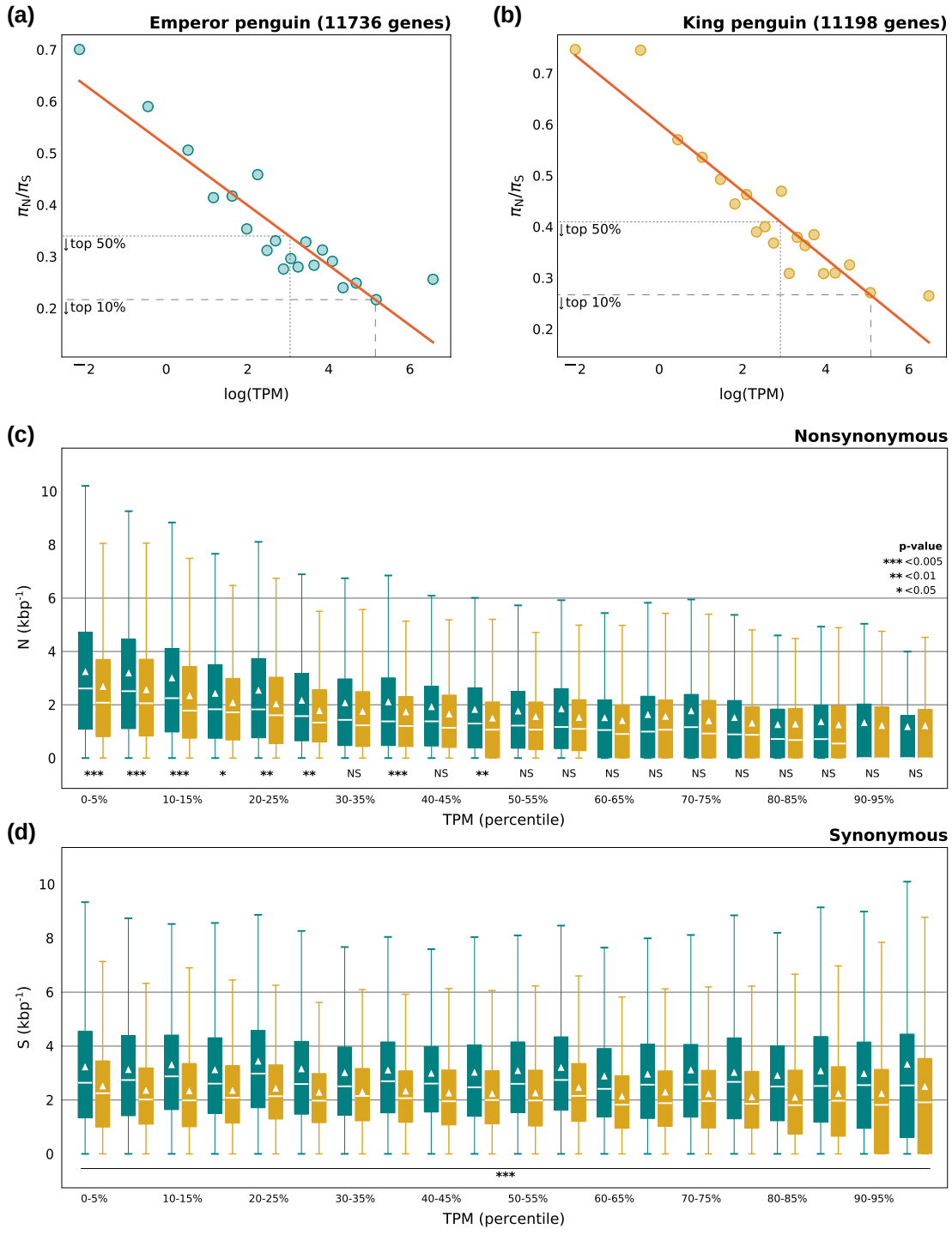

Fig. 3. Increasing purifying selection with gene expression in Emperor (teal) and King (gold) penguins. Estimates of π_N/π_S (a, b), and average number of nonsynonymous (c) and synonymous (d) segregating variants (normalized per 1,000 bp of coding sequence) in genes binned by 5% percentiles of expression rate (normalized as TPM). Slope of the linear regression (Emperor penguin: $\chi = -0.058$, $R^2 = 0.848$; King penguin: $\chi = -0.066$, $R^2 = 0.877$) is shown as a solid red line, and values of π_N/π_S for the top 50% (small dashes) and 10% (large dashes) of the most highly expressed genes are indicated by a dashed gray line in a) and b). Median (solid white line) and mean (white triangle) are shown in each boxplot in c) and d). Linear regression of the number of variants as a function of gene expression, as log(TPM), was fitted independently for each penguin species and each type of segregating variant (regression lines not shown): Emperor penguin synonymous ($\chi = -0.01722$, $R^2 = 0.07682$) and nonsynonymous ($\chi = -0.28309$; $R^2 = 0.9427$) and King penguin synonymous ($\chi = -0.006407$; $R^2 = 0.01836$) and nonsynonymous ($\chi = -0.18945$; $R^2 = 0.9212$). Statistical significance for the difference in the distribution of synonymous and nonsynonymous variants per percentile between the two species is shown (KS test: * < 0.05, ** < 0.01, *** < 0.001) in b) and c).

assumptions that proteins are selected for their conformational stability or for protein–protein interaction, these results suggest that both the effects of effective population size and gene expression can be considered together in integrated models of evolution. However, they should be assessed more thoroughly, by comparing more population sizes. On a different note, even if gene expression has been suggested to be one of the causes of nonneutrality in synonymous variants in yeast (Shen et al. 2022), or that codon usage bias is more intense in highly expressed genes (Frumkin et al. 2018), gene expression rate does

not appear to perturb synonymous variation in our datasets from two vertebrate species (Fig. 3d).

Purifying selection more efficiently prevents nonsynonymous segregating variants from increasing in frequency in genes with higher expression rate

The derived allele frequency spectrum of nonsynonymous variants with expression rate higher than 0.3 CPM is depleted in medium–high-frequency categories, while there is no difference in the derived allele frequency spectrum of synonymous variants across the whole expression range in both species: the fitted smoothing splines, which describe the pattern of the derived allele frequencies along the count classes, significantly differ between highly and lowly expressed genes only for nonsynonymous variants (Fig. 4). As predicted by theoretical models (Nielsen 2005; Lawrie and Petrov 2014) and empirical results (Paape et al. 2018; Han et al. 2019), the skewness to the left of the nonsynonymous variants site frequency spectrum demonstrates more intense purifying selection in genes with higher expression rate. Changing the arbitrary threshold to discriminate between low and high expression rate, or using more than two categories of expression rate (low: < 0.3 CPM, medium: 0.3 to 2 CPM, high: > 2 CPM) does not change the observed pattern (supplementary fig. S10, Supplementary Material online). The pattern holds when all nonsynonymous and synonymous variants are used in the allele frequency spectrum estimate (supplementary fig. S10, Supplementary Material online) as well as when one nonsynonymous and one synonymous variant are randomly sampled from each gene (Fig. 4), thus excluding the possibility that few genes with many variants (i.e. pseudoreplication) drive our observation.

Purifying selection in the top 10% of highly expressed genes largely exceeds the effect of 100,000 individuals' effective population size

In simulated populations, under either Wright–Fisher or more realistic non-Wright–Fisher models, median uncorrected p_N/p_S across genes declines from 1.8 to 0.9, while population size increases from 1,000 to 100,000 individuals (Fig. 5). Such values of p_N/p_S are much higher than the values observed in penguin populations for genes in the top 50% or top 10% of expression rate (Fig. 5 and supplementary fig. S6, Supplementary Material online). In these models, the effect of population size on purifying selection was explored by simulating a set of realistic values for mutation and recombination rate, synonymous to nonsynonymous ratio, selection and dominance coefficient distributions, coding sequence length, and gene numbers. In particular, new mutations were given a selection and dominance coefficient (h-mix) based on a commonly used fitness effects distribution (Kim et al. 2017; Kyriazis et al. 2021), where most of the mutations are weakly deleterious. To reproduce p_N/p_S values as those observed in highly expressed genes in both penguin species, we designed a more extreme selection scenario: all nonsynonymous mutations appearing in a gene were given a fixed selection coefficient of −0.1, −0.01, or −0.001 (100 replicated genes per selection coefficient) and a dominance coefficient derived from the hs relationship (Henn et al. 2016). In realistic non-Wright–Fisher models with a selection coefficient of −0.01, p_N/p_S decreases below 0.4 (Fig. 5), while a selection coefficient of −0.1 results in p_N/p_S below 0.3, as observed in genes in the top 50% and 10% of expression rate, respectively (Fig. 3a and supplementary fig. S6, Supplementary Material online). Such high selection coefficients are then expected to be effective even when the population size is small (i.e. $s >> 1/N_e$, per $n = 1{,}000$), thus buffering the effects of changing

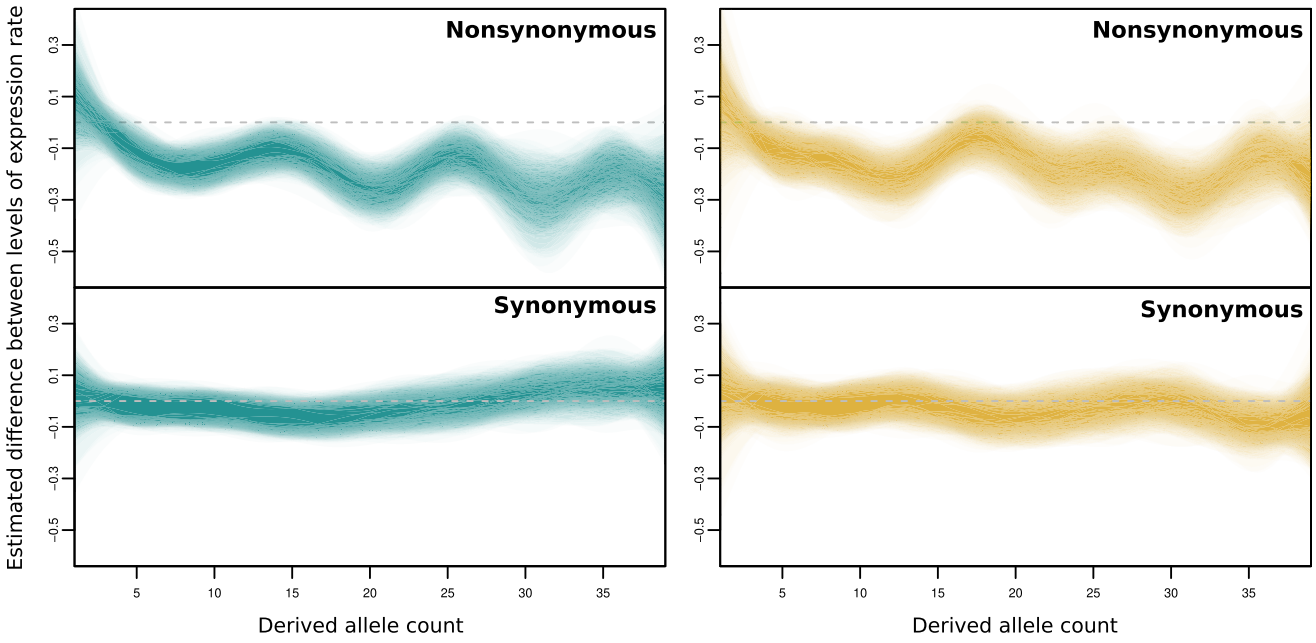

Fig. 4. Site frequency spectra significantly differ for nonsynonymous variants between highly and lowly expressed genes. Overlapping 95% CIs of the differences between fitted smoothing splines of the derived allele frequency spectra of nonsynonymous and synonymous variants at low (<0.3 CPM) and high (>0.3 CPM) expression levels in Emperor (teal) and King (gold) penguins. Each shaded polygon represents the 95% CI of the pairwise difference between the high- and the low-expression smoothing splines obtained by fitting a generalized additive mixed-effects model on a random subsample of the data (for a total of 100 subsamples). Intervals overlapping with the dashed gray line at zero indicate no significant difference between the two smoothing splines.

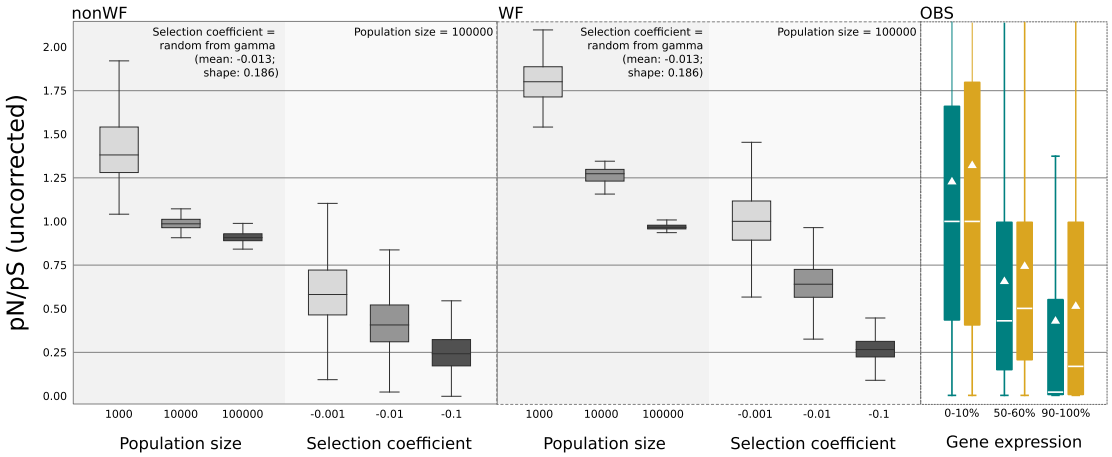

Fig. 5. Population size and gene-specific extreme selection coefficients explain low observed *pN/pS* values in simulations. Distribution of uncorrected p_N/p_S (2.3 expected under neutrality) across 1,000 genes simulated under non-WrightFisher (non-WF, left, solid border) and WrightFisher (WF, right, dashed border) models with effective population size from 1,000 to 100,000 (darker gray background) and across 100 genes with selection coefficient from −0.001 to −0.1 (lighter gray background). Note that the dominance coefficient is set according to the *h-mix* or *hs* models in simulations testing different population sizes or selection coefficients, respectively. More efficient purifying selection in nonWF models, where effective population size tends to be lower than in WF models, can be explained by the fact that, in such models, individuals with high fitness can survive and reproduce for multiple generations. OBS: observed uncorrected p_N/p_S in three selected bins of gene expression (TPM percentile as in Fig. 3); Emperor penguin: teal; King penguin: gold; solid white line: median; white triangle: mean.

Table 1 Expression rate by predicted fitness effect

Predicted fitness effect	LOWsyn		MDR		HIGH	
Species	Emperor	King	Emperor	King	Emperor	King
Total number of variants	73501	57088	47183	41352	934	840
Expression rate of all variants	1.50 (5.02)	1.30 (4.56)	0.78 (2.85)	0.78 (2.86)	0.24 (1.97)	0.14 (0.88)
Number of fixed differences	1846	3500	1166	2229	16	44
Expression rate of fixed differences	1.41 (4.12)	1.22 (4.14)	1.40 (6.31)	0.93 (3.39)	0.08 (0.54)	0.02 (0.57)
Average count of derived alleles in segregating variants	6.17	6.25	4.75	4.73	5.14	5.48

LOWsyn, low effect and synonymous; MDR, moderate effect; HIGH, high effect. Expression rate is shown as average (and SD) *Z*-normalized CPM.

population size as it was also suggested for the vast majority of X-linked genes in *Drosophila* (Andolfatto et al. 2011). However, we also observe more variance in simulations with smaller population sizes (supplementary fig. S11, Supplementary Material online). Of note, the effects of strongly deleterious mutations are difficult to estimate as few mutations actually segregate in a sample of haplotypes (Charmouh et al. 2023; Zeng et al. 2024). However, consistently with our results, joint inference of mean population-size-scaled selection coefficients (N_e *s*) in some humans and *Drosophila* samples have been estimated between 2,500 and 14,000 (Keightley and Eyre-Walker 2007).

Gene expression can be used as a proxy of the distribution of gene selection coefficients in natural populations of nonmodel species

Variants with highly deleterious effects on individual fitness are expected to be immediately lost in natural populations. Consistent with this expectation, the highly deleterious variants (<1,000 HIGH effect SNPs per species) predicted by SNPeff (Cingolani et al. 2012) in each penguin population show a much lower average expression level than weakly deleterious (MODERATE effect) and nearly-neutral (LOW effect and synonymous) variants (Table 1). This observation means that HIGH effect variants are mainly present in lowly expressed genes with limited impact on fitness. Expression level is even lower in the very few fixed differences with HIGH effect (Table 1), thus supporting our hypothesis. As site-specific expression of highly deleterious variants (mainly start/stop codons loss/gain and splice acceptor/donor variants) could be biassed in mature mRNA sequencing, we also estimated the expression of highly deleterious variants as the expression of the gene they belong to. Even applying a rather conservative test, the expression of genes with predicted highly deleterious variants is on average three times lower than all genes (KS test *P*-value = 0.00095) in the King penguin and slightly lower, even if not significant, in the Emperor penguin. As previously suggested for the distribution of dominance coefficients in a model plant species (Huber et al. 2018), gene expression should be taken into account when using predictions of fitness effects and, more generally, when using such predictions to calculate the genetic load in populations of conservation concern (Bertorelle et al. 2022). In fact, predicted highly deleterious variants could be in lowly expressed genes, thus with little contribution to individual or population fitness.

Conclusion

Overall, our study provides evidence that gene expression can have a major impact on purifying selection in natural populations and to a higher extent than 100,000 individuals' population size for highly expressed genes. About half of the genes in

a genome, which are likely responsible for basic cellular and molecular functions (Boyle et al. 2017), are under a strong selective constraint preventing deleterious sequence changes even when population size declines to about 1,000 individuals. Indeed, for selection to be effective the ratio $N_e s$ has to be >1. Selection coefficients on the top 10% of the expressed genes could be so high (s up to 0.1) to prevent fixation of deleterious mutations even at smaller population size (<100 individuals). However, below this order of magnitude, random effects would necessarily prevail in the proteins' evolutionary trajectories. Together with the effect on cell and organism function, the probability of fixation of a new mutation appears as also determined by an additional component of selection which scales with gene expression. One of the possible explanations is that novel mutations could result in increased cellular toxicity due to protein misfolding or protein–protein misinteractions and that such toxicity increases with the amount of the produced protein (Zhang and Yang 2015), which is approximated by gene expression. However, gene expression is correlated with gene multifunctionality, centrality in regulatory and genetic interactions network, and essentiality for cell basal functioning (Pritykin et al. 2015). Therefore, the observed relationship between gene expression and purifying selection could be a covariate of other gene features. On a practical note, gene expression should be integrated in estimates of gene selection coefficients, which are notoriously difficult to study in natural populations of nonmodel species (Huber et al. 2017), or when accurate estimates of π_N/π_S are hampered by low levels of segregating variants as in very small populations (Benazzo et al. 2017). Gene expression data are easier to collect than selection coefficients and are usually highly conserved across closely related species (Fig. 2d), so that they can be used to refine estimates of genetic load (Bertorelle et al. 2022) in natural populations of conservation concern.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors thank Fabrizio Mafessoni, Paolo Gratton, Robin Cristofari, and Matthew Hahn for helpful comments and suggestions on the analyses and Federica Pirri for RNA data production. This study was approved by the French ethics committee (last: APAFIS#29338-2020070210516365) and the French Polar Environmental Committee, and permits to handle animals and access breeding sites were delivered by the "Terres Australes et Antarctiques Françaises" (TAAF). The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Norway Regional Health Authorities. Bioinformatic analyses were performed on the HPC clusters of the Department of Life and Environmental Sciences ("HappyComputingDiSVA"), Marche Polytechnic University, and of the Department of Life Sciences and Biotechnology, University of Ferrara. A preprint version of this article has been peer-reviewed and recommended by PCI Evolutionary Biology (https://doi.org/10.24072/pci.evolbiol.100705).

Author Contributions

E.T. designed the study and secured funding together with G.B., extracted the DNA samples, analyzed the genomic and transcriptomic data, and wrote the manuscript. P.M. and F.G. built, performed, and analyzed the genomic simulations and wrote the relevant section in the Methods. T.L. performed the regression analyses between purifying selection, expression rate, and population size and wrote the relevant text (section 3) in the Extended Methods. M.G. performed statistical analyses. N.C.S. contributed to the pilot study leading to this work. F.A.N.F., L.A., J.F.O., J.P., and G.B. discussed the results and contributed to the manuscript C.L.B. coordinated the project logistics and the samples collection and the associated funding, discussed the results, and contributed to the manuscript. All authors contributed to the development and finalization of the manuscript.

Funding

The study was supported by PNRA_16 00164 ("Programma Nazionale di Ricerca in Antartide." Bando PNRA 5 aprile 2016, no. 651—Linea B "Genomica degli adattamenti estremi alla vita in Antartide"; PI: E.T.), by the Institut Polaire Français Paul-Emile Victor (IPEV) within the framework of the Program 137-ANTAVIA (PI: C.L.B.), by the Centre Scientifique de Monaco with additional support from the LIA-647 and RTPI-NUTRESS (CSM/CNRS-UNISTRA), and by the Centre National de la Recherche Scientifique (CNRS) through the Programme Zone Atelier de Recherches sur l'Environnement Antarctique et Subantarctique (ZATA) and the long-term Studies in Ecology and Evolution (SEE-Life) programme.

Data Availability

Genomic and transcriptomic raw reads are publicly available at NCBI and ENA database with project accession numbers PRJNA1099460 and PRJEB64484, respectively. The filtered SNPs dataset is available here: 10.5281/zenodo.10688854. Bioinformatic scripts are available here: github.com/emitruc/ExpressionLoad; github.com/ThibaultLatrille/PenguinExpression; and github.com/PiergiorgioMassa/penguin_gene_expression_slimulations.

References

Akashi H, Osada N, Ohta T. Weak selection and protein evolution. *Genetics*. 2012:192(1):15–31. https://doi.org/10.1534/genetics.112.140178.

Andolfatto P, Wong KM, Bachtrog D. Effective population size and the efficacy of selection on the X chromosomes of two closely related Drosophila species. *Genome Biol Evol*. 2011:3:114–128. https://doi.org/10.1093/gbe/evq086.

Bédard C, Cisneros AF, Jordan D, Landry CR. Correlation between protein abundance and sequence conservation: what do recent experiments say? *Curr Opin Genet Dev*. 2022:77:101984. https://doi.org/10.1016/j.gde.2022.101984.

Benazzo A, Trucchi E, Cahill JA, Maisano Delser P, Mona S, Fumagalli M, Bunnefeld L, Cornetti L, Ghirotto S, Girardi M, *et al*. Survival and divergence in a small group: the extraordinary genomic history of the endangered Apennine brown bear stragglers. *Proc Natl Acad Sci U S A*. 2017:114(45):E9589–E9597. https://doi.org/10.1073/pnas.1707279114.

Bertorelle G, Raffini F, Bosse M, Bortoluzzi C, Iannucci A, Trucchi E, Morales HE, van Oosterhout C. Genetic load: genomic estimates

and applications in non-model animals. *Nat Rev Genet*. 2022:23(8):492–503. https://doi.org/10.1038/s41576-022-00448-x.
Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell*. 2017:169(7):1177–1186. https://doi.org/10.1016/j.cell.2017.05.038.
Carneiro M, Albert FW, Melo-Ferreira J, Galtier N, Gayral P, Blanco-Aguiar JA, Villafuerte R, Nachman MW, Ferrand N. Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. *Mol Biol Evol*. 2012:29(7):1837–1849. https://doi.org/10.1093/molbev/mss025.
Charlesworth B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*. 2009:10(3):195–205. https://doi.org/10.1038/nrg2526.
Charmouh AP, Bocedi G, Hartfield M. Inferring the distributions of fitness effects and proportions of strongly deleterious mutations. *G3 (Bethesda)*. 2023:13(9):jkad140. https://doi.org/10.1093/g3journal/jkad140.
Chen J, Glémin S, Lascoux M. Genetic diversity and the efficacy of purifying selection across plant and animal species. *Mol Biol Evol*. 2017:34(6):1417–1428. https://doi.org/10.1093/molbev/msx088.
Cheng C, Kirkpatrick M. Molecular evolution and the decline of purifying selection with age. *Nat Commun*. 2021:12(1):1–10. https://doi.org/10.1038/s41467-020-20314-w.
Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012:6(2):80–92. https://doi.org/10.4161/fly.19695.
Cole TL, Zhou C, Fang M, Pan H, Ksepka DT, Fiddaman SR, Emerling CA, Thomas DB, Bi X, Fang Q, *et al*. Genomic insights into the secondary aquatic transition of penguins. *Nat Commun*. 2022:13(1):3912. https://doi.org/10.1038/s41467-022-31508-9.
Cristofari R, Bertorelle G, Ancel A, Benazzo A, Le Maho Y, Ponganis PJ, Stenseth NC, Trathan PN, Whittington JD, Zanetti E, *et al*. Full circumpolar migration ensures evolutionary unity in the emperor penguin. *Nat Commun*. 2016:7(1):11842. https://doi.org/10.1038/ncomms11842.
Cristofari R, Liu X, Bonadonna F, Cherel Y, Pistorius P, Le Maho Y, Raybaud V, Stenseth NC, Le Bohec C, Trucchi E. Climate-driven range shifts of the king penguin in a fragmented ecosystem. *Nat Clim Chang*. 2018:8(3):245–251. https://doi.org/10.1038/s41558-018-0084-2.
Duret L, Mouchiroud D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol*. 2000:17(1):68–070. https://doi.org/10.1093/oxfordjournals.molbev.a026239.
Figuet E, Nabholz B, Bonneau M, Mas Carrio E, Nadachowska-Brzyska K, Ellegren H, Galtier N. Life history traits, protein evolution, and the nearly neutral theory in amniotes. *Mol Biol Evol*. 2016:33(6):1517–1527. https://doi.org/10.1093/molbev/msw033.
Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y. Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proc Natl Acad Sci U S A*. 2018:115(21):E4940–E4949. https://doi.org/10.1073/pnas.1719375115.
Galtier N. Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genet*. 2016:12(1):e1005774. https://doi.org/10.1371/journal.pgen.1005774.
Han S, Andrés AM, Marques-Bonet T, Kuhlwilm M. Genetic variation in Pan species is shaped by demographic history and harbors lineage-specific functions. *Genome Biol Evol*. 2019:11(4):1178–1191. https://doi.org/10.1093/gbe/evz047.
Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, Martin AR, Musharoff S, Cann H, Snyder MP, *et al*. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc Natl Acad Sci U S A*. 2016:113(4):E440–E449. https://doi.org/10.1073/pnas.1510805112.
Hodgins KA, Yeaman S, Nurkowski KA, Rieseberg LH, Aitken SN. Expression divergence is correlated with sequence evolution but not positive selection in conifers. *Mol Biol Evol*. 2016:33(6):1502–1516. https://doi.org/10.1093/molbev/msw032.
Huber CD, Durvasula A, Hancock AM, Lohmueller KE. Gene expression drives the evolution of dominance. *Nat Commun*. 2018:9(1):2750. https://doi.org/10.1038/s41467-018-05281-7.
Huber CD, Kim BY, Marsden CD, Lohmueller KE. Determining the factors driving selective effects of new nonsynonymous mutations. *Proc Natl Acad Sci U S A*. 2017:114(17):4465–4470. https://doi.org/10.1073/pnas.1619508114.
Josephs EB, Wright SI, Stinchcombe JR, Schoen DJ. The relationship between selection, network connectivity, and regulatory variation within a population of *Capsella grandiflora*. *Genome Biol Evol*. 2017:9(4):1099–1109. https://doi.org/10.1093/gbe/evx068.
Keightley PD, Eyre-Walker A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*. 2007:177(4):2251–2261. https://doi.org/10.1534/genetics.107.080663.
Kim BY, Huber CD, Lohmueller KE. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*. 2017:206(1):345–361. https://doi.org/10.1534/genetics.116.197145.
Kofler R, Schlötterer C. A guide for the design of evolve and resequencing studies. *Mol Biol Evol*. 2014:31(2):474–483. https://doi.org/10.1093/molbev/mst221.
Kyriazis CC, Wayne RK, Lohmueller KE. Strongly deleterious mutations are a primary determinant of extinction risk due to inbreeding depression. *Evol Lett*. 2021:5(1):33–47. https://doi.org/10.1002/evl3.209.
Latrille T, Lartillot N. Quantifying the impact of changes in effective population size and expression level on the rate of coding sequence evolution. *Theor Popul Biol*. 2021:142:57–66. https://doi.org/10.1016/j.tpb.2021.09.005.
Lawrie DS, Petrov DA. Comparative population genomics: power and principles for the inference of functionality. *Trends Genet*. 2014:30(4):133–139. https://doi.org/10.1016/j.tig.2014.02.002.
Nielsen R. Molecular signatures of natural selection. *Annu Rev Genet*. 2005:39(1):197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420.
Paape T, Briskine RV, Halstead-Nussloch G, Lischer HEL, Shimizu-Inatsugi R, Hatakeyama M, Tanaka K, Nishiyama T, Sabirov R, Sese J, *et al*. Patterns of polymorphism and selection in the subgenomes of the allopolyploid *Arabidopsis kamchatica*. *Nat Commun*. 2018:9(1):3909. https://doi.org/10.1038/s41467-018-06108-1.
Pál C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. *Genetics*. 2001:158(2):927–931. https://doi.org/10.1093/genetics/158.2.927.
Park C, Chen X, Yang JR, Zhang J. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. 2013:110(8):E678–E686. https://doi.org/10.1073/pnas.1218066110.
Pirri F, Ometto L, Fuselli S, Fernandes FA, Ancona L, Perta N, Di Marino D, Le Bohec C, Zane L, Trucchi E. Selection-driven adaptation to the extreme Antarctic environment in the emperor penguin. *Heredity (Edinb)*. 2022:129(6):317–326. https://doi.org/10.1038/s41437-022-00564-8.
Pritykin Y, Ghersi D, Singh M. Genome-wide detection and analysis of multifunctional genes. *PLoS Comput Biol*. 2015:11(10):e1004467. https://doi.org/10.1371/journal.pcbi.1004467.
Shen X, Song S, Li C, Zhang J. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature*. 2022:606(7915):725–731. https://doi.org/10.1038/s41586-022-04823-w.
Shibai A, Kotani H, Sakata N, Furusawa C, Tsuru S. Purifying selection enduringly acts on the sequence evolution of highly expressed proteins in Escherichia coli. *G3 (Bethesda)*. 2022:12(11):jkac235. https://doi.org/10.1093/g3journal/jkac235.
Slotte T, Bataillon T, Hansen TT, St. Onge K, Wright SI, Schierup MH. Genomic determinants of protein evolution and polymorphism in *Arabidopsis*. *Genome Biol Evol*. 2011:3:1210–1219. https://doi.org/10.1093/gbe/evr094.

Trucchi E, Gratton P, Whittington JD, Cristofari R, Le Maho Y, Stenseth NC, Le Bohec C. King penguin demography since the last glaciation inferred from genome-wide data. *Proc R Soc Lond B Biol Sci*. 2014:281(1787):20140528. https://doi.org/10.1098/rspb.2014.0528.

Vianna JA, Fernandes FAN, Frugone MJ, Figueiró HV, Pertierra LR, Noll D, Bi K, Wang-Claypool CY, Lowther A, Parker P, *et al*. Genome-wide analyses reveal drivers of penguin diversification. *Proc Natl Acad Sci U S A*. 2020:117(36):22303–22310. https://doi.org/10.1073/pnas.2006659117.

Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of Capsella grandiflora. *PLoS Genet*. 2014:10(9):e1004622. https://doi.org/10.1371/journal.pgen.1004622.

Wu Z, Cai X, Zhang X, Liu Y, Tian G-B, Yang J-R, Chen X. Expression level is a major modifier of the fitness landscape of a protein coding gene. *Nat Ecol Evol*. 2022:6(1):103–115. https://doi.org/10.1038/s41559-021-01578-x.

Yang JR, Liao BY, Zhuang SM, Zhang J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc Natl Acad Sci U S A*. 2012:109(14):E831–E840. https://doi.org/10.1073/pnas.1117408109.

Zeng T, Spence JP, Mostafavi H, Pritchard JK. Bayesian estimation of gene constraint from an evolutionary model with gene features. *Nat Genet*. 2024:56(8):1632–1643. https://doi.org/10.1038/s41588-024-01820-9.

Zhang J, Yang JR. Determinants of the rate of protein sequence evolution. *Nat Rev Genet*. 2015:16(7):409–420. doi:10.1038/nrg3950.