Phylogenetic analysis reveals how selection and mutation shape the coevolution of mRNA and protein abundances

Alexander L. Cope[1,2,*], Joshua G. Schraiber[3,*] & Matt Pennell[3,4,†]

[1]*Department of Genetics, Rutgers University, USA*
[2]*Robert Woods Johnson Medical School, Rutgers University, USA*
[3]*Department of Quantitative and Computational Biology, University of Southern California, USA*
[4]*Department of Biological Sciences, University of Southern California, USA*
[*]These authors contributed equally
[†]Corresponding author: mpennell@usc.edu

Summary The regulatory mechanisms that shape mRNA and protein abundances are intensely studied. Much less is known about the evolutionary processes that shape the relationship between these two levels of gene expression. To disentangle the contributions of mutational and selective processes, we derive a novel phylogenetic model and fit it to multi-species data from mammalian skin tissue. We find that over macroevolutionary time: 1) there has been strong stabilizing selection on protein abundances; 2) mutations impacting mRNA abundances have minimal influence on protein abundances; 3) mRNA abundances are under selection to track protein abundances, and 4) mRNA abundances adapt more quickly than protein abundances due to increased mutational opportunity. We find additional support for these findings by comparing gene-specific parameter estimates from our model to human functional genomic data. More broadly, our new phylogenetic approach provides a foundation for testing hypotheses about the processes that led to divergence in gene expression.

Introduction

Evolutionary divergence in gene expression is a major contributor to phenotypic divergence [1–4]. Accordingly, in the past decades, many studies leveraged high-throughput DNA microarrays and RNAseq to investigate how evolutionary processes shape patterns of mRNA abundance [4–11]. However, it is clear that mRNA abundances provide an incomplete picture of the evolution of gene expression and that we also need to consider evolution at the level of protein abundances [12–16]; there is no reason to assume that these are exchangeable quantities. Indeed, a fundamental question in molecular biology is understanding how various regulators and associated molecular mechanisms — from transcription to protein degradation — act in concert to produce the necessary amount of protein of a given gene for a cell to function [17–19]. The precise degree of correlation between steady-state abundances of mRNA and protein abundances across different genes, and how this correlations varies across organisms, is highly debated owing to various technical and statistical issues [20, 21]. But critically, even if we were able to unambiguously estimate their relationship, it is not clear what, if anything, this tells us about the evolutionary processes that shape the relationship between these two levels of gene expression.

In addition to the degree of correlation, there are some important lines of evidence that can help us understand the evolutionary causes of this relationship. First, genetic variants associated with changes in expression of mRNA (i.e., eQTLs) or proteins (i.e., pQTLs) are often distinct [22–24]. Second, comparative studies revealed that protein abundances are generally more conserved than mRNA abundances [12, 13, 25]. Third, there is some evidence (though it is disputed on statistical grounds; [26]) of an evolutionary buffering effect in which evolutionary changes at one level of regulation can be offset by subsequent changes to another level of regulation, ultimately leading to little change in protein abundances [14–16, 27]. This could be do to off-setting changes in transcription and translation, as has been documented in yeasts [14, 15], or between transcription and post-translational regulation (i.e., degradation), as has been found in primates [16, 22]. We note that the mechanisms involved in such evolutionary buffering may or may not be distinct from the buffering mechanisms that keep protein levels relatively stable despite transcriptional bursts and other causes of "noise" in gene expression [28–34].

However, these on their own are not enough to develop a comprehensive explanation for observed patterns of gene expression. An intuitive appealing explanation of the evidence presented above is that of "compensatory evolution", in which protein abundances, which are "closer" to the phenotype than mRNA abundances, are generally under stabilizing selection (i.e., there is an evolutionary optimum value for these levels produced from a cell); regulatory mutations that push proteins away from this optimum can fix if they are compensated by other types of regulatory mutations that move the protein abundances back. However, this model is inconsistent with the changes to transcription and translation that occurred during a long-term experimental evolution study [35]. The model also implicitly predicts that mRNA abundances should show interspecific divergences that look like those produced by neutrality [36], which is also inconsistent with abundant comparative evidence that mRNA abundances are highly constrained [4, 37–39]. And critically, being a purely verbal model, the compensatory evolution explanation neither makes any predictions for how

much mRNA and protein abundances should diverge from another (i.e., at what observed level of correlation would this explanation no longer be valid?) nor the relative importance of various evolutionary or regulatory processes in shaping patterns of gene expression across lineages.

What is needed to gain a richer understanding of these processes is a statistical approach, rooted in first evolutionary and molecular principles, that can be fit to different realizations of the evolutionary process (i.e., comparative data). To this end, we derived a set of novel phylogenetic models and Bayesian Markov Chain Monte Carlo machinery for fitting this model to data. By comparing the fit of alternative models to matched multi-species transcriptome and proteome datasets from mammalian skin cell samples and by examining the values of estimated parameters, our model reveals new insights into the sources of evolutionary divergence in gene expression. To ensure that our model is capturing the key processes, we validate the predictions of our model using independent functional genomic data from humans, a species not included in the dataset we analyzed.

Results

Mathematical modeling

We briefly describe the derivation of our novel phylogenetic modeling framework; the full details of the various models and their derivation are available in the Supplementary Materials. We start with the basic premise that there are two possible types of mutations. First, there are "mRNA mutations" that influence the steady-state mRNA abundance and, in turn, impact the downstream protein steady-state abundances; this includes both mutations that influence transcription and those that impact mRNA degradation, which are distinct in terms of the molecular mechanism, but will have similar impacts on the evolutionary dynamics. We can describe the effect of mRNA mutations with a simple mutational model. Let R denote the log(steady-state mRNA abundance) and P denote the log(steady-state protein abundance). Then, mRNA mutations occur at rate μ_R per genome per generation and result in the following changes to mRNA and protein abundances:

$$R \rightarrow R' = R + \delta R \tag{1}$$

$$P \rightarrow P' = P + c\delta R \tag{2}$$

where c is a constant that relates changes in mRNA to changes in protein, and δR is a random variable with mean 0 and variance σ_R^2 that quantifies the distribution of mutational effects on mRNA abundance. The parameter c can be interpreted as the degree to which mRNA mutations are buffered ($|c| < 1$) or amplified ($|c| > 1$) at the protein level.

Second, there are “protein mutations” that impact protein abundance but do not impact mRNA abundance (i.e., mutations that influence translation initiation or protein degradation); importantly, these do not directly influence steady-state mRNA abundances. In our model, protein mutations occur at rate μ_P per

genome per generation:

$$R \to R' = R \tag{3}$$

$$P \to P' = P + \delta P. \tag{4}$$

where δP is a random variable with mean 0 and variance σ_P^2 that quantifies the distribution of mutational effects on protein abundance that do not impact mRNA abundance.

To model the fitness of given mRNA and protein abundances, we wish to capture the idea that there may be a cost to having mRNA and protein abundances that are "mismatched", for example, if too much mRNA is produced to achieve a given protein abundance, there may be metabolic costs to the excess transcription [40, 41]. Thus, we assume that the fitness of a given R is a function of P, and vice versa. Using a Gaussian fitness function, as is common in evolutionary quantitative genetics [42], we set

$$w(R, P) \propto \exp\left(-\frac{(P - \theta_P - a_P R)^2}{2V_P} - \frac{(R - \theta_R - a_R P)^2}{2V_R}\right). \tag{5}$$

The first term of the fitness function describes the fitness contribution of the mRNA and the second, the fitness consequences of proteins. In each case, there is an optimum value (θ_R and θ_P) independent of that of the other and a fitness cost of being off this optimum. Importantly, there are also terms that describe how strong selection is for the mRNA abundances to match to protein abundances a_R and as well as for the protein abundances to match the mRNA abundances a_P. The terms V_P and V_R are the width of the fitness function; note that smaller V_i indicates stronger selection, as the fitness function becomes more peaked. (Here and elsewhere, we use the subscript i to indicate cases where we are referring to both the mRNA and protein analogs of a parameter.)

In the Supplementary Materials, we incorporate the above mutational and fitness models into a strong selection, weak mutation framework in which deleterious or beneficial mutations are either expunged from or fixed in diploid population of size N before another one arises [43]. This allows us to derive a pair of coupled stochastic differential equations describing the joint evolution of steady-state mRNA and protein abundances subject to mutation and selection. In deriving the model, we find that compound parameters describing the combined effects of mutation and selection are important. First, $\alpha = 2N\mu_R\frac{\sigma_R^2}{V_R}$ describes the rate of adaptation of the mRNA. Second, $\beta = 2N\mu_P\frac{\sigma_P^2}{V_P}$ describes the rate of adaptation of the protein. Note that in both cases, the rate of adaptation is a function of mutational target size (as indicated by the μ_i), mutational effect size (as indicated by the σ_i), and the strength of selection (as indicated by the V_i). Mutation and genetic drift also influence evolution through the parameters $\tau_i^2 = \mu_i\sigma_i^2$.

The most general model, in which all the parameters can take arbitrary values, can generate complex feedback loops between R and P. Such a model is both computationally intractable and very difficult to interpret. We therefore focused on three limiting cases of our general model. These cases represent different ends of the spectrum of possible evolutionary scenarios. By comparing these simplified versions, we can gain insights into the general dynamics of the mRNA-protein relationship, and by looking at the parameter

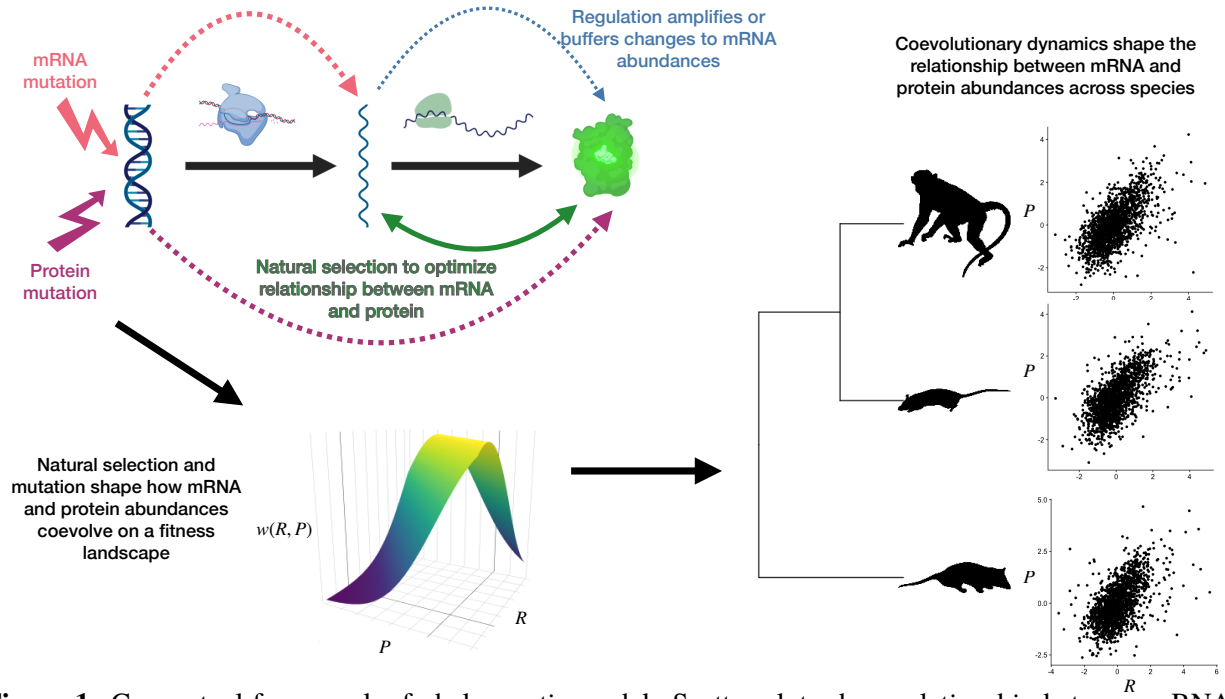

Figure 1: Conceptual framework of phylogenetic model. Scatter plots show relationship between mRNA and protein abundances are shown for *Macaca mulatta*, *Rattus norvegicus*, and *Monodelphis domestica* from [25]. Icons for DNA, RNA polymerase, mRNA, ribosome, and protein were obtained from BioRender.com. Images of species were taken from Phylopic(`https://www.phylopic.org/`).

values estimated under these different models, assess the relative importance of different processes. In all of these models, we describe the dynamics using a diffusion process, where $\mathbf{B}_t$ is uncorrelated two-dimensional Brownian motion.

In the first model, we assume that mRNA mutations play an important role on macroevolutionary timescales, and that natural selection acting on mRNA abundances is substantially stronger than natural selection acting on those of proteins. We refer to this model as "mRNA-dominant evolution" and in the Supplementary Materials, we show that this model can be written

$$\begin{bmatrix} dR_t \\ dP_t \end{bmatrix} = -\begin{bmatrix} \alpha & 0 \\ \alpha c - \beta a_P & \beta \end{bmatrix} \left(\begin{bmatrix} R_t \\ P_t \end{bmatrix} - \begin{bmatrix} \theta_R \\ a_P \theta_R + \theta_P \end{bmatrix} \right) dt + \begin{bmatrix} \tau_R & 0 \\ c\tau_R & \tau_P \end{bmatrix} d\mathbf{B}_t. \tag{6}$$

In the second model, mRNA mutations have a negligible impact on evolutionary changes to protein abundances, with evolutionary changes to mRNA abundances largely responding to changes to protein abundances. Here, we set $c = 0$, this would imply the effects of mRNA mutations are buffered, suggesting the existence of regulatory mechanisms downstream of mRNA abundances to leave protein abundances largely unaltered. We refer to this model as "protein-driven evolution" and show that this can be written as:

$$\begin{bmatrix} dR_t \\ dP_t \end{bmatrix} = -\begin{bmatrix} \alpha & -a_R \alpha \\ 0 & \beta \end{bmatrix} \left(\begin{bmatrix} R_t \\ P_t \end{bmatrix} - \begin{bmatrix} a_R \theta_P + \theta_R \\ \theta_P \end{bmatrix} \right) dt + \begin{bmatrix} \tau_R & 0 \\ 0 & \tau_P \end{bmatrix} d\mathbf{B}_t. \tag{7}$$

Last, as a null model, we considered that the evolutionary dynamics of mRNA and protein abundance are uncoupled. We refer to this model as "independent evolution". Then, the model can be written

$$\begin{bmatrix} dR_t \\ dP_t \end{bmatrix} = -\begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} \left(\begin{bmatrix} R_t \\ P_t \end{bmatrix} - \begin{bmatrix} \theta_R \\ \theta_P \end{bmatrix} \right) dt + \begin{bmatrix} \tau_R & 0 \\ 0 & \tau_P \end{bmatrix} d\mathbf{B}_t. \tag{8}$$

Conveniently, all of these equations turn out to be coupled Ornstein-Uhlenbeck (OU) processes. There are known algorithms to compute the likelihood of observing a set of values at the tips of the phylogeny given the tree structure and generalized coupled OU models [44]. Univariate OU processes (in which, like our null model, adaptation in one variable does not depend on any other measured variable) have been widely applied to describe the evolution of mRNA counts along a phylogeny [4, 37, 38, 45–47] but to our knowledge, this is the first application of a coupled OU processes to study gene expression evolution. It is also notable that we derived OU processes using a strong selection weak mutation modeling framework whereas OU processes have conventionally been derived by assuming constant additive genetic variance [42, 48]; we suspect that this may be a useful general result for evolutionary theory. In order to fit special cases of our model using a Bayesian approach, we developed a custom Markov Chain Monte Carlo (MCMC) machinery for estimating the parameters using code from the `PCMBase` [49] and `LaplacesDemons` libraries. Using extensive simulations (see Supplementary Materials), we validated that our estimation procedure was able to recover the generating parameters under a range of conditions and that standard Bayesian model selection techniques — specifically, the Deviance Information Criterion (DIC) [50] — are able to correctly identify the generating scenario from the posterior distributions.

Testing evolutionary hypotheses using mammalian skin fibroblast expression data

We analyzed a recently published dataset [25] of mRNA and protein abundances measured for ten species of mammalian skin samples. Critically, these were all measured by the same group of researchers, using standardized experimental protocols to minimize the impact of technical variation. We note that the published dataset also included measurements from an 11th species — humans — but these were from cell line data and, likely for this reason, showed aberrant patterns of gene expression compared to rest of the species and so were excluded from our analyses (see Supplementart Material). The phylogenetic tree of these ten species and the correlations between mRNA and protein levels are shown in Figure 2A; the pairwise similarity as a function of divergence time is shown in Figure 2B. Ten species is a large sample size for paired transcriptomic and proteomic data but it is nonetheless small relative to the number of parameters we sought to estimate. To overcome this limitation, we assumed that the steady-state mRNA and protein abundances for each of the 1,641 orthologous genes in the dataset was an independent outcome of the same co-evolutionary process. As such, we estimated unique steady-state optima, θ_R and θ_P, for each gene but set the other parameters to be equal across genes. Although there is certainly variation in the co-evolutionary dynamics across genes, we demonstrated using simulations that our model parameter estimates closely match the mean values across genes (see Supplementary Materials).

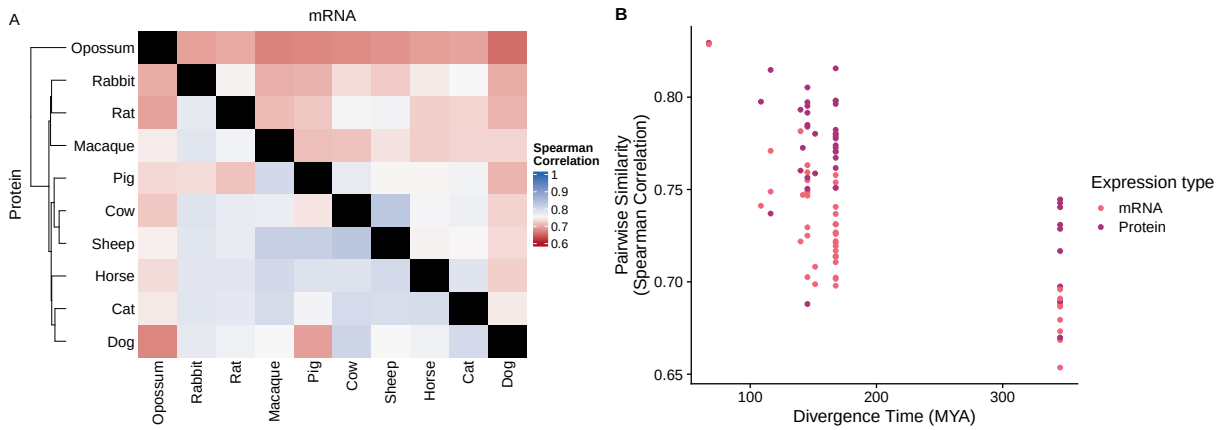

Figure 2: (A) Heatmap showing across-species pairwise comparisons (measured as the Spearman rank correlation) of mRNA (upper triangle) and protein abundances (lower triangle). The phylogenetic tree of the 10 species is included on the lower triangle. (B) Pairwise correlations of mRNA and protein abundance as a function of species divergence time.

Overall, we find overwhelming support for the protein-driven evolution model (Figure 3A: the DIC value for the protein-driven model is 4,201 points below the mRNA-driven evolution model and 5,129 points below the independent model; a typical rule of thumb is that a 10 point difference is sufficient to strongly support one model over another [51]).

Examining the parameter values estimated from the preferred protein-driven evolution model, we find that the two parameters describing the rate of adaptation, α and β parameter have posterior means of 0.022 (95% Credible Interval (CI): 0.019–0.024) and 0.016 (95% CI: 0.014–0.018), respectively (Figure 3B). These estimates correspond to a mean "phylogenetic half-life" [48] of 31 Million Years (MY) (95% CI: 28–34 MY) for the mRNA abundances and 41 MY (95% CI: 36–46 MY) for the protein abundances. This means that it takes longer for the proteins to adapt to changes in their optima than it does for mRNA and is consistent with previous observations that protein abundances are more highly similar between species than mRNA abundances [12, 13, 25]. Parameter estimates are highly consistent across independent MCMC chains starting from different points in parameter space (see Supplementary Materials).

As noted above, the a_R parameter in the protein-driven model represents the linear strength of selection for the mRNA abundances to match the protein abundances; we estimated this parameter to be 0.742 (95% of CI: 0.702–0.782, Figure 3C). The observation that a_R is less than but relatively close to 1 suggests there is strong selection for the mRNA levels to match the protein levels but that there is a substantial evolutionary lag for the alignment to equilibrate following selection for changes to the protein levels.

The posterior means for the stochastic (i.e., drift) parameters τ_R and τ_P are estimated to be very similar (τ_R is 0.088 (95% CI: 0.084 – 0.092) and τ_P is 0.086 (95% CI: 0.082 – 0.091) (Figure 3D). Taken with the above estimates, this implies that while the selective dynamics are very different between mRNA and protein abundances, the role of drift in shaping their abundances is essentially the same for both.

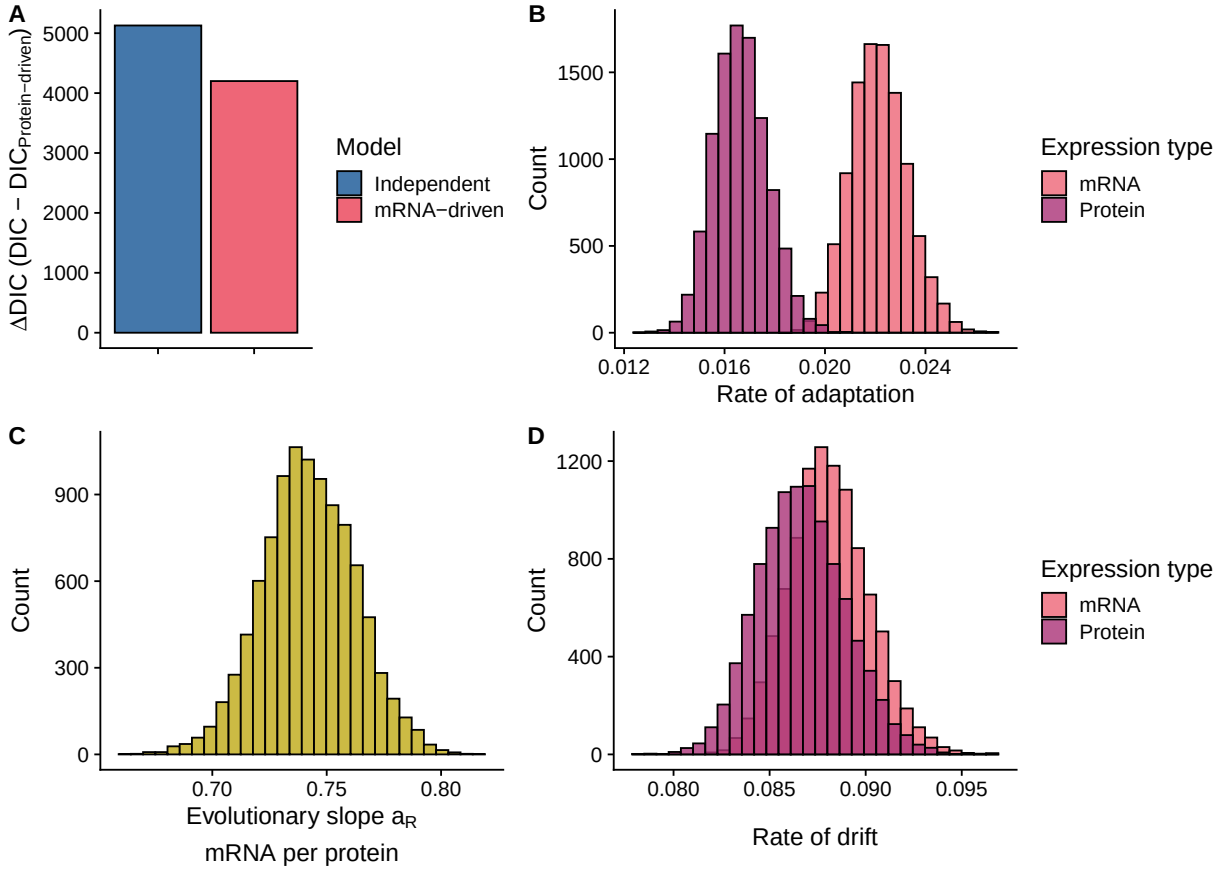

Figure 3: Comparisons of models and posterior distributions of model parameters from protein-driven model fit. (A) Comparison of DIC scores for mRNA-driven and Independent models relative to the protein-driven model, the overall best model. (B) mRNA and protein abundance rate of adaptation parameters α and β, respectively. (C) Evolutionary slope a_R relating the coevolution of mRNA and protein abundances. (D) mRNA and protein abundance rate of drift parameters τ_R and τ_P, respectively.

Validating predictions of evolutionary model using independent functional genomic data from humans

For the sake of tractability, we had to make some strong, simplifying assumptions about the complex evolutionary processes that led to patterns of gene expression divergence. As such, there may be alternative interpretations of the parameter values we estimated (i.e., if we left out some key process, its effect may be included in the estimates of parameters we did include; [4]). However, our interpretations do make some clear predictions, which we tested using complementary functional genomic data from human studies — these data provide critical out-of-sample tests of our model. As the protein-driven model was better at explaining the co-evolutionary dynamics of mRNA and protein abundances overall, we will focus on applying this model to subsets of genes.

First, various lines of evidence indicate natural selection is stronger on highly expressed genes [52–57]. If our interpretation of our model fits is correct, we predict adaptation parameters α and β are correlated with expression level. To avoid the circularity of stratifying our analyses by mean expression level, which is itself a parameter estimated in the model, we binned genes into deciles based on an integrated measure of protein abundances in humans taken from PaxDB 5.0 [58]; we emphasize again that humans were not included in our analyses. We then re-fit the preferred protein-driven model to the binned data (i.e., co-evolutionary parameters were assumed to be constant within a bin but each bin was estimated independently). We found that genes that were more highly expressed tended to have stronger rate of adaptation parameters for both mRNA and protein abundances, except for the highest expression categories in the case of proteins, in which the direction of the trend reversed (Figure 4A). (We do not calculate *P*-values as the sample size is a property of the binning strategy; we divided the data into gene expression deciles to ensure that we could still estimate parameters within each set of binned genes.) To ensure this was not an anomaly, we grouped the genes into deciles based on the mean expression level across all species in the phylogenetic dataset and re-ran the analysis; we discovered a nearly identical trend (see Supplementary Material).

We inferred that $\alpha > \beta$ for both the entire dataset and many of the expression bins, which we interpret to mean that mRNA abundances adapt more rapidly than protein abundances either due to increased mutational input or due to stronger selection acting on mRNA abundances. To our knowledge, there are no experimental results that directly support this. However, the number of *cis*-regulatory elements could be used to test the general proposition that the estimated strength of selection from the phylogenetic models is directly related to mutational input (assuming that the mutational target size is predictive of mutational input). Indeed, mutation accumulation experiments suggest that genes with larger mutational target sizes generally have larger mutational variances [59]. To do this, we obtained predicted enhancer information from the EnhancerAtlas2.0 database [60] to estimate the number of active enhancers per gene in human keratinocyte cells, the primary cell type found in the epidermis. To control for confounding due to the absolute value of gene expression (which we showed does have an effect above), we calculated the mean number of enhancers per gene for each expression bin and compared this average to the estimated parameter value of each bin. Consistent with our interpretation, genes with more enhancers in human keratinocyte cells show a faster rate of adaptation at the mRNA levels (Spearman rank correlation $R = 0.94$, Figure 4B) but not the protein level.

Furthermore, our interpretation of the $\alpha > \beta$ results also predict that the stochastic component of evolution, τ_i, should increase with increased mutational input. As noted above, we observed that the rate of stochastic rate of evolution was nearly identical between mRNA and protein abundances (Figure 3D). However, this apparent similarity is confounded by gene expression variation: once we stratified by gene expression bin, we find that the number of enhancers is highly correlated with the difference in the stochastic rates of mRNA compared to proteins (Spearman rank correlation $R = 0.78$; Figure 4C). As an additional verification of both of the results involving enhancer number (i.e., Figures 4B,C), we performed the same analysis using enhancers identified in mouse kidney cells and fruit fly larval BG3-c2 larval cell lines [60]. While using the mouse kidney enhancers produced results similar to those observed using the human dataset (see Supplementary Material), there was no relationship between parameter estimates and the number of enhancers estimated for *Drosophila melanogaster* (see Supplementary Material). We anticipate that the explanatory power of our gene-specific estimates should erode over deep evolutionary time as the regulatory landscape evolves.

As an alternative measure of mutational input, we examined eQTLs in humans using data from GTEx (version 8) [62, 63]. We calculated the mean absolute magnitude (i.e., the absolute value of the effect) of eQTLs for each gene across all available human datasets in GTEx, which includes multiple tissue-types. As with enhancers, we then calculated the mean eQTL across all genes within each decile bin based on the independent measure of gene expression. We predicted that because the rate of adaptation depends on the variance of mutational effects, genes with a greater rate of adaptation might be expected to have larger magnitude eQTLs. Surprisingly, we see the opposite trend, as genes with larger eQTL evolve under weaker selection (Figure 4D). We speculate this is because eQTLs need to be both common enough and have large enough effects to be detected using genome-wide association study methods. Thus, eQTL for genes under strong selection must have small effects, while large effect eQTL must impact genes under weak selection. Our observation here dovetails with the recent result of Mostafavi and colleagues [64] who argued that there is little overlap between genes with eQTL and genes associated with disease due to this trade-off between effect size and strength of selection.

Finally, we tested whether genes whose functional roles make them likely to be under strong purifying selection indeed have stronger selection terms estimated from our model (Figure 4E). Using annotations from [61], we found that, as predicted by our protein-driven model, the rate of adaptation for protein abundances was significantly stronger in essential genes (on natural scale, $\beta = 0.048$, 95% CI: 0.027 – 0.071) versus non-essential genes (on natural scale, $\beta = 0.015$, 95% CI: 0.013 – 0.017). There was no difference between the selection on mRNA between the essential (on natural scale, $\alpha = 0.0229$, 95% CI: 0.015 – 0.030) and non-essential genes (on natural scale, $\alpha = 0.0228$, 95% CI: 0.020 – 0.025), consistent with our inference that natural selection on protein levels is the primary driver of gene expression evolution.

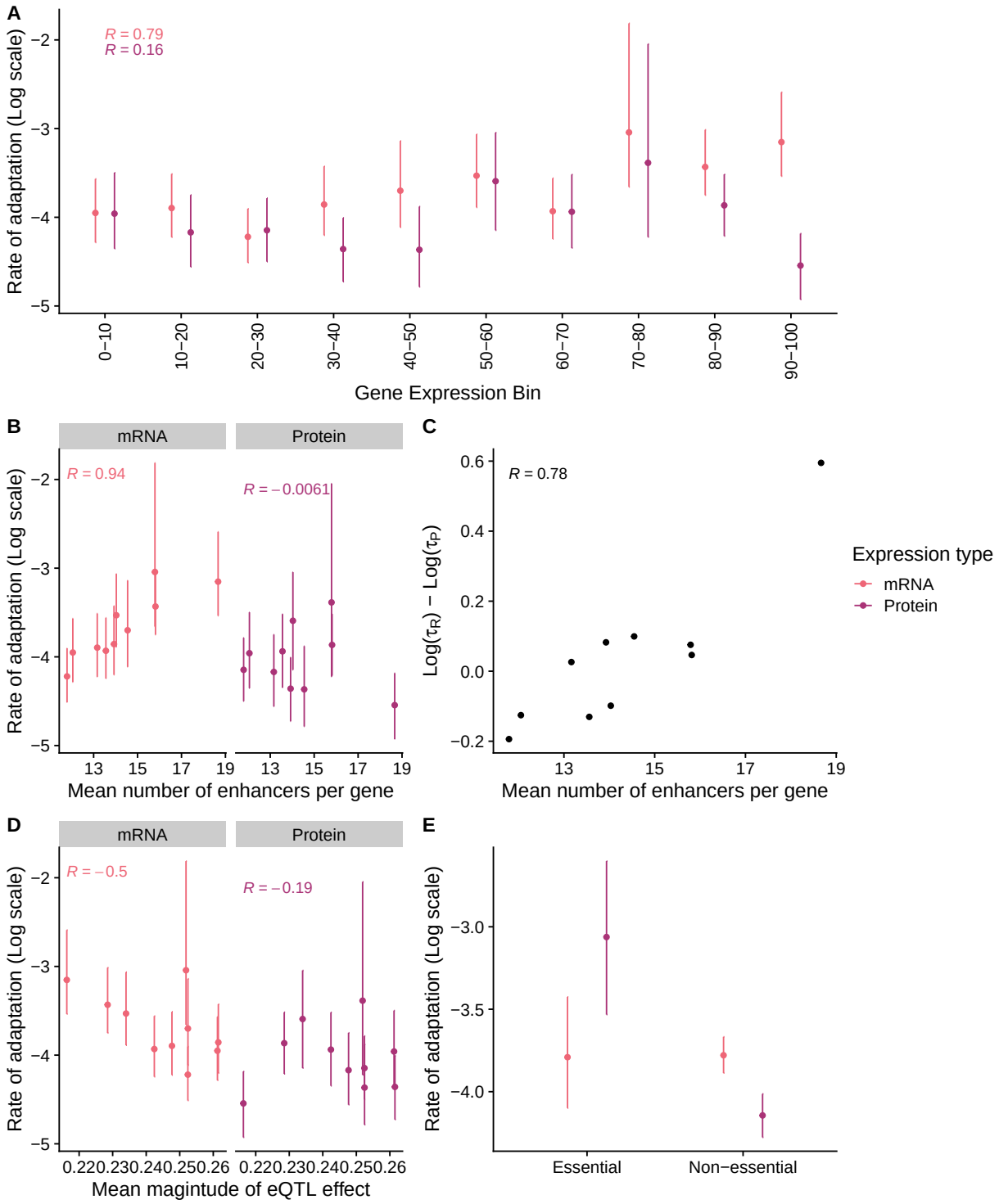

Figure 4: Validation of model parameter estimates using out-of-sample human functional genomics data. Fits were based on protein-driven model. All error bars represent 95% credible intervals. Spearman rank correlations R are reported. (A) Rate of adaptation parameter α (mRNA) and β (protein) across gene expression decile bins. Deciles were determined using the integrated protein abundance data for humans from PaxDB. (B) The relationship between the mean number of enhancers per gene and the rate of adaptation for mRNA and protein abundances. Rate of adaptation parameters are taken from the decile bins based on gene expression in (A). (C) The relationship between differences in the rate of drift (log-scale) and the mean number of enhancers per gene. Rate of drift parameters are taken from the decile bins based on gene expression in (A). (D) Similar to (B), but using the mean absolute magnitude of eQTL effects in humans obtained from GTEx. (E) Rate of adaptation parameters across "Essential" and "Non-essential" genes in humans *in vitro* cell lines as previously defined [61].

Discussion

Here, we applied a novel phylogenetic comparative model to quantify the evolutionary dynamics of gene expression across a 10-species mammalian lineage. Our work makes a critical step forward in studying gene expression evolution on macroevolutionary timescales: unlike previous work, we focus on the co-evolutionary dynamics of mRNA and protein abundances. In doing so, we gained new insights into the evolutionary and mechanistic processes shaping gene expression evolution. Our overarching finding is the very strong support for the protein-driven evolution model of gene expression evolution. From this support and the estimates of the specific parameters of the model, we find support for four major claims.

Strong selection to maintain protein abundances

First, we find evidence for strong stabilizing selection on protein abundances; as predicted, this is even stronger for essential genes than on non-essential genes, and stronger for highly expressed genes compared to low expressed ones. Moreover, we find that genes with large-effect eQTL show evidence of weaker selection than those with smaller effects. This is consistent with only being able to ascertain large-effect eQTL in genes evolving under weak selection, while genes subject to strong selection can only tolerate small-effect eQTL segregating at frequencies that would enable their detection in a genome-wide association study [64].

Regulatory mechanisms that impact protein abundances independent of mRNA abundances play a significant role in gene expression evolution

Second, mRNA mutations (e.g., those that impact transcription) have a negligible impact on protein abundances over long evolutionary timescales. Even though we rejected the mRNA-driven model in favor of the protein-driven model, the posterior mean of c – which represents the downstream effect of mutations that alter mRNA abundances on protein abundances – was 0.141 (95% CI: 0.079 – 0.197), indicating only a small proportion of the effect of mRNA mutations carries over to protein abundances. This suggests cells have various mechanisms to minimize the impact of variation in mRNA abundances on protein production [18, 19, 30–32, 62, 65–69]

Our conclusions may, at first glance, also appear to be inconsistent with a large body of research in evolutionary developmental biology showing the evolution of enhancers to be responsible for innovation and adaptation across a diverse number of lineages [2, 70]. However, one possible way to reconcile these findings is suggested by our model and borne out by our analysis of functional genomic data: the rate of adaptation of mRNA abundance is higher than the rate of adaptation for protein abundance, due primarily to larger mutational input in mRNA abundances. Thus, as protein abundances shift, mutations that impact mRNA abundance can be rapidly fixed to canalize those shifts to protein abundance. Further modeling and mechanistic work may be able to test this model. Moreover, we emphasize that what we are estimating is averaged over many types of mutations and over many millions of years and does not contradict any of the

notable discoveries of enhancer-driven evolution [71–74].

Strong selection to maintain correlations between mRNA and protein abundances

Third, we find evidence for strong selection to maintain mRNA abundances that closely match their corresponding protein abundances. In other words, the degree of matching we and others have observed between mRNA and protein abundances is consistent with selection for relatively tight correlations. Our best-fit model contrasts with a cellular biology model in which mRNA and protein abundances are correlated primarily due to how the machinery of the cell produces proteins from mRNA transcripts. Unlike those studies, our goal here was not to predict protein abundance from mRNA abundance, but to understand the evolutionary processes shaping the correlation between these abundances that represent the tuning of non-independent regulatory mechanisms. This highlights the value of building the relevant mechanisms into the statistical model itself because it is not obvious from looking at any particular level of correlation which mechanisms need to be invoked.

Why might natural selection act to tune steady-state mRNA abundance to match the required steady-state protein abundances if the protein abundance is truly the functional unit? One possibility relates to noise in gene expression. Recent work suggested that discrepancies between transcription and translation to achieve the same protein abundance could either increase noise or increase the cost of gene expression, with the former appearing generally more tolerable [41]. We note that a_R increases with gene expression, suggesting natural selection related to the fine-tuning of mRNA and protein abundances is stronger in highly expressed genes (see Supplementary Material).

Interestingly, while our parameter estimates of the evolutionary slope a_R relating mRNA and protein abundances across all genes is relatively close to 1, the posterior distribution does not contain 1. Taking the inverse of a_R gives a value greater than 1 and is in units of protein per mRNA, similar to estimates of translation efficiency often calculated from ribosome profiling and RNA-seq data [14, 15, 26]. This could be interpreted as evidence that the optimal relationship [75] between mRNA and protein abundances is one in which downstream regulation (e.g., translation) amplifies, rather than buffers, steady-state abundances. Although not a comparative study, previous work in *S. cerevisiae* concluded that – after accounting for noise using both structural equation modeling and ranged major axis regression — the relationship between mRNA and protein abundance supports amplification of mRNA abundances due to coevolution between transcription and translation [21]. Our results generally support this conclusion.

mRNA abundances evolve quickly due to greater mutational input

Last, the fit of our model implies that the per-generational mutational input (the change in the phenotype owing to mutation) is larger for mRNA abundances than for protein abundances. To our knowledge, this has not been directly tested before. Recent work used simulations to show that greater divergences in transcription (relative to translation) between yeast paralogs could be explained by a "mutation bias," towards

transcriptional changes [76]. This can be assessed using mutation-accumulation experiments [77], in which the experimental design reduces the efficacy of selection such that most mutations are observed. Such experiments allow for directly characterizing the distribution of phenotypic effects [78, 79].

One proxy for mutational input is the mutational target size: the greater the number of regulatory targets, the greater impact we expect mutations to have in shaping gene expression evolution just by chance. Previous mutation accumulation experiments in yeasts revealed mutational variance V_M correlated with the number of *cis*- (promoters, enhancers) and *trans*- (transcription factors) regulatory elements [78]. Functional genomics studies have revealed the complexity of gene regulation, including how regulation can shift over the course of development, making determining the exact mutational target size of a gene for mRNA and protein abundances difficult. Aside from the challenges of identifying and linking causal genetic variants impacting gene expression evolution, eQTLs are generally more comprehensive than pQTLs due to the relative ease and availability of high-throughput sequencing compared to mass spectrometry [4], although the number of pQTL datasets is growing [22–24]. Despite these challenges, we found that the number of enhancers of different genes predicts the evolutionary dynamics in the same way we would predict if regulatory target size were a proxy for mutational variance. Aside from increased mutational input, increasing the number of regulatory elements is hypothesized to improve the robustness of gene expression to mutations [80–83]. Taking these results in the context of our work suggests a positive correlation between the evolvability and robustness of gene expression evolution, particularly mRNA abundances [84].

Limitations of current approach

There are some important caveats to our analyses. First, as noted throughout, we made some strong assumptions when deriving our evolutionary model as well as some additional statistical assumptions (e.g., different genes share evolutionary dynamics). We also chose to look at a limited set of special cases; there are other biologically plausible configurations that we did not consider here. We made these assumptions for the sake of mathematical and computational tractability but we suspect that a combination of some clever statistical approaches and the larger comparative transcriptomic/proteomic datasets that we anticipate are on the horizon may enable researchers to develop more sophisticated elaborations of our general framework. Second, as with essentially all phylogenetic comparative studies of gene expression, we are only looking at a single measure (steady-state abundance) for each gene and these measures are from an isolated environment in a highly controlled experiment. This is naturally a limited perspective of gene expression [85]. Third and also in line with the majority of phylogenetic comparative studies of gene expression, our work is focused on a set of highly-conserved one-to-one orthologs. This will naturally bias our results towards genes under relatively strong evolutionary constraints and precludes studying the impact of gene duplication on gene expression evolution [76, 86, 87]. Finally, our model assumes gene expression evolves independently across genes, ignoring both shared regulatory elements and correlated selection (i.e., related to functional constraints) [88, 89]. Despite these limitations, we were able to validate the central predictions of our model and its interpretations with independent human functional genomic data — to our knowledge, no other study of the phylogenetic dynamics of gene expression has done — which suggests that, to a first-order approximation,

our model captured at least some of the major features of the evolutionary dynamics.

Concluding remarks

Numerous studies have used phylogenetic models to investigate the evolutionary processes that shaped variation in mRNA abundance counts among species [4, 37, 38, 90, 91]. In this paper, we take this work a big step forward and use phylogenetic models to infer the processes that have shaped the co-evolution between mRNA and protein levels. Doing so required us to derive novel, mechanistic models from first principles to capture the biochemical and evolutionary relationships between these variables. We suggest that this general modeling framework can serve as the foundation for further investigations into the role of complex processes in shaping patterns of expression. And remarkably, we show that the evolutionary parameters we estimated from these models are consistent with within-species functional genomic information. There is a long-standing debate as to how well macroevolutionary divergences can be explained in terms of microevolutionary processes [92–94]; our results imply that more integrative approaches, that combine within and between-species data [*sensu* [95]] can provide richer insights into gene expression evolution than investigating either in isolation.

Data availability

No new data were generated for this study. The publicly-available data used in this study are available via the references provided in this manuscript. R notebooks and R scripts for recreating our analyses can be found at `https://github.com/phylo-lab-usc/rna-protein-coevolution`.

Acknowledgements

We thank Rex Jiang for comments on the manuscript, as well as the Pennell, Edge, and Mooney lab groups for helpful discussion of this work. Fabio Machado and Josef Uyeda provided advice on our implementation of the MCMC. Tony Zheng, Jeffrey Spence, Hakhamanesh Mostafavi, and Jonathan Pritchard provided curated data on gene function. Gunter Wagner provided the phylogenetic tree associated with the expression data analyzed here. This work was supported by the NIH-funded Rutgers INSPIRE IRACDA Postdoctoral Program (grant #GM093854 to ALC) and NIH grant R35GM151348 to MP.

References

[1] Wray, G. A. (2007) The evolutionary significance of cis-regulatory mutations.

[2] Carroll, S. B. (2008) Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution.

[3] Albert, F. W & Kruglyak, L. (2015) *Nature Reviews Genetics 2015 16:4* **16**, 197–212.

[4] Price, P. D, Droguett, D. H. P, Taylor, J. A, Kim, D. W, Place, E. S, Rogers, T. F, Mank, J. E, Cooney, C. R, & Wright, A. E. (2022) *Nature Ecology & Evolution* **6**, 1035–1045.

[5] Khaitovich, P, Weiss, G, Lachmann, M, Hellmann, I, Enard, W, Muetzel, B, Wirkner, U, Ansorge, W, & Pääbo, S. (2004) *PLoS Biology* **2**, e132.

[6] Lemos, B, Meiklejohn, C. D, Cáceres, M, & Hartl, D. L. (2005) *Evolution* **59**, 126–137.

[7] Khaitovich, P, Enard, W, Lachmann, M, & Pääbo, S. (2006) *Nature Reviews Genetics 2006 7:9* **7**, 693–702.

[8] Blekhman, R, Oshlack, A, Chabot, A. E, Smyth, G. K, & Gilad, Y. (2008) *PLOS Genetics* **4**, e1000271.

[9] Wittkopp, P. J & Kalay, G. (2012) Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence.

[10] Romero, I. G, Ruvinsky, I, & Gilad, Y. (2012) *Nature Reviews Genetics 2012 13:7* **13**, 505–516.

[11] Hill, M. S, Zande, P. V, & Wittkopp, P. J. (2021) *Nature Reviews Genetics* **22**, 203–215.

[12] Laurent, J. M, Vogel, C, Kwon, T, Craig, S. A, Boutz, D. R, Huse, H. K, Nozue, K, Walia, H, Whiteley, M, Ronald, P. C, & Marcotte, E. M. (2010) *PROTEOMICS* **10**, 4209–4212.

[13] Khan, Z, Ford, M. J, Cusanovich, D. A, Mitrano, A, Pritchard, J. K, & Gilad, Y. (2013) *Science (New York, N.Y.)* **342**, 1100–4.

[14] Artieri, C. G & Fraser, H. B. (2014) *Genome Research* **24**, 411–421.

[15] McManus, C. J, May, G. E, Spealman, P, & Shteyman, A. (2014) *Genome Research* **24**, 422–430.

[16] Wang, S. H, Hsiao, J, Khan, Z, & Pritchard, J. K. (2018) *Genome Biology* **19**.

[17] Vogel, C & Marcotte, E. M. (2012) *Nature Reviews Genetics* **13**.

[18] Liu, Y, Beyer, A, & Aebersold, R. (2016) *Cell* **165**, 535–550.

[19] Buccitelli, C & Selbach, M. (2020) *Nature Reviews Genetics* **21**, 630–644.

[20] Li, J. J, Bickel, P. J, & Biggin, M. D. (2014) *PeerJ* **2014**, 1–26.

[21] Csárdi, G, Franks, A, Choi, D. S, Airoldi, E. M, & Drummond, D. A. (2015) *PLOS Genetics* **11**, e1005206.

[22] Battle, A, Khan, Z, Wang, S. H, Mitrano, A, Ford, M. J, Pritchard, J. K, & Gilad, Y. (2015) *Science* **347**, 664–667.

[23] Chick, J. M, Munger, S. C, Simecek, P, Huttlin, E. L, Choi, K, Gatti, D. M, Raghupathy, N, Svenson, K. L, Churchill, G. A, & Gygi, S. P. (2016) *Nature* **534**, 500–505.

[24] Teyssonnière, E. M, Trébulle, P, Muenzner, J, Loegler, V, Ludwig, D, Amari, F, Mülleder, M, Friedrich, A, Hou, J, Ralser, M, & Schacherer, J. (2024) *Proceedings of the National Academy of Sciences of the United States of America* **121**, e2319211121.

[25] Ba, Q, Hei, Y, Dighe, A, Li, W, Maziarz, J, Pak, I, Wang, S, Wagner, G. P, & Liu, Y. (2022) *Science Advances* **8**, 756.

[26] Albert, F. W, Muzzey, D, Weissman, J. S, & Kruglyak, L. (2014) *PLOS Genetics* **10**, e1004692.

[27] Wang, Z. Y, Leushkin, E, Liechti, A, Ovchinnikova, S, Mößinger, K, Brüning, T, Rummel, C, Grützner, F, Cardoso-Moreira, M, Janich, P, Gatfield, D, Diagouraga, B, de Massy, B, Gill, M. E, Peters, A. H, Anders, S, & Kaessmann, H. (2020) *Nature* **588**, 642–647.

[28] Lareau, L. F, Inada, M, Green, R. E, Wengrod, J. C, & Brenner, S. E. (2007) *Nature* **446**.

[29] Battich, N, Stoeger, T, & Pelkmans, L. (2015) *Cell* **163**, 1596–1610.

[30] Goncalves, E, Fragoulis, A, Garcia-Alonso, L, Cramer, T, Saez-Rodriguez, J, & Beltrao, P. (2017) *Cell Systems* **5**.

[31] Kustatscher, G, Grabowski, P, & Rappsilber, J. (2017) *Molecular Systems Biology* **13**, 937.

[32] Grabowski, P, Kustatscher, G, & Rappsilber, J. (2018) *Molecular and Cellular Proteomics* **17**, 2082–2090.

[33] Popovic, D, Koch, B, Kueblbeck, M, Ellenberg, J, & Pelkmans, L. (2018) *Cell Systems* **7**.

[34] Waszak, S. M, Delaneau, O, Gschwind, A. R, Kilpinen, H, Raghav, S. K, Witwicki, R. M, Orioli, A, Wiederkehr, M, Panousis, N. I, Yurovsky, A, Romano-Palumbo, L, Planchon, A, Bielser, D, Padioleau, I, Udin, G, Thurnheer, S, Hacker, D, Hernandez, N, Reymond, A, Deplancke, B, & Dermitzakis, E. T. (2015) *Cell* **162**.

[35] Favate, J. S, Liang, S, Cope, A. L, Yadavalli, S. S, & Shah, P. (2022) *eLife* **11**.

[36] Jiang, D, Cope, A. L, Zhang, J, & Pennell, M. (2023) *Molecular Biology and Evolution* **40**.

[37] Brawand, D, Soumillon, M, Necsulea, A, Julien, P, Csárdi, G, Harrigan, P, Weier, M, Liechti, A, Aximu-Petri, A, Kircher, M, Albert, F. W, Zeller, U, Khaitovich, P, Grützner, F, Bergmann, S, Nielsen, R, Pääbo, S, & Kaessmann, H. (2011) *Nature* **478**, 343–348.

[38] Chen, J, Swofford, R, Johnson, J, Cummings, B. B, Rogel, N, Lindblad-Toh, K, Haerty, W, Palma, F. D, & Regev, A. (2019) *Genome Research* **29**, 53–63.

[39] Dimayacyac, J. R, Wu, S, Jiang, D, & Pennell, M. (2023) *Genome Biology and Evolution* **15**.

[40] Wagner, A. (2005) *Molecular Biology and Evolution* **22**, 1365–1374.

[41] Hausser, J, Mayo, A, Keren, L, & Alon, U. (2019) *Nature Communications 2019 10:1* **10**, 1–15.

[42] Lande, R. (1976) *Evolution* **30**, 314–334.

[43] Gillespie, J. H. (1983) *The American Naturalist* **121**, 691–708.

[44] Bartoszek, K, Pienaar, J, Mostad, P, Andersson, S, & Hansen, T. F. (2012) *Journal of Theoretical Biology* **314**, 204–215.

[45] Rohlfs, R. V, Harrigan, P, & Nielsen, R. (2014) *Molecular Biology and Evolution* **31**, 201–211.

[46] Rohlfs, R. V & Nielsen, R. (2015) *Systematic biology* **64**, 695–708.

[47] Vaishnav, E. D, de Boer, C. G, Molinet, J, Yassour, M, Fan, L, Adiconis, X, Thompson, D. A, Levin, J. Z, Cubillos, F. A, & Regev, A. (2022) *Nature* **603**, 455–463.

[48] Hansen, T. F. (1997) *Evolution* **51**, 1341–1351.

[49] Mitov, V, Bartoszek, K, Asimomitis, G, & Stadler, T. (2020) *Theoretical Population Biology* **131**.

[50] Spiegelhalter, D. J, Best, N. G, Carlin, B. P, & Linde, A. V. D. (2002) *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **64**, 583–616.

[51] Burnham, K. P & Anderson, D. R. (2004) *Sociological Methods & Research* **33**, 261–304.

[52] Drummond, D. A, Raval, A, & Wilke, C. O. (2006) *Molecular Biology and Evolution* **23**, 327–337.

[53] Drummond, D. A & Wilke, C. O. (2008) *Cell* **134**, 341–352.

[54] Shah, P & Gilchrist, M. (2011) *Proceedings of the National Academy of Sciences of the United States of America* **108**, 10231–10236.

[55] Yang, J.-R, Liao, B.-Y, Zhuang, S.-M, & Zhang, J. (2012) *Proceedings of the National Academy of Sciences of the United States of America* **109**, E831–40.

[56] Cope, A. L & Gilchrist, M. A. (2022) *BMC Genomics* **23**, 408.

[57] Zhang, J & Yang, J.-R. (2015) *Nature Reviews Genetics* **16**, 409–420.

[58] Huang, Q, Szklarczyk, D, Wang, M, Simonovic, M, & von Mering, C. (2023) *Molecular and Cellular Proteomics* **22**, 100640.

[59] Landis, M. J, Schraiber, J. G, & Liang, M. (2013) *Systematic biology* **62**, 193–204.

[60] Gao, T & Qian, J. (2020) *Nucleic Acids Research* **48**, D58–D64.

[61] Zeng, T, Spence, J. P, Mostafavi, H, & Pritchard, J. K. (2023) *bioRxiv*.

[62] Carithers, L. J, Ardlie, K, Barcus, M, Branton, P. A, Britton, A, Buia, S. A, Compton, C. C, Deluca, D. S, Peter-Demchok, J, Gelfand, E. T, Guan, P, Korzeniewski, G. E, Lockhart, N. C, Rabiner, C. A, Rao, A. K, Robinson, K. L, Roche, N. V, Sawyer, S. J, Segrè, A. V, Shive, C. E, Smith, A. M, Sobin, L. H, Undale, A. H, Valentino, K. M, Vaught, J, Young, T. R, Moore, H. M, Barker, L, Basile, M, Battle, A, Boyer, J, Bradbury, D, Bridge, J. P, Brown, A, Burges, R, Choi, C, Colantuoni, D, Cox, N, Dermitzakis, E. T, Derr, L. K, Dinsmore, M. J, Erickson, K, Fleming, J, Flutre, T, Foster, B. A, Gamazon, E. R, Getz, G, Gillard, B. M, Guigó, R, Hambright, K. W, Hariharan, P, Hasz, R, Im, H. K, Jewell, S, Karasik, E, Kellis, M, Kheradpour, P, Koester, S, Koller, D, Konkashbaev, A, Lappalainen, T, Little, R, Liu, J, Lo, E, Lonsdale, J. T, Lu, C, MacArthur, D. G, Magazine, H, Maller, J. B, Marcus, Y, Mash, D. C, McCarthy, M. I, McLean, J, Mestichelli, B, Miklos, M, Monlong, J, Mosavel, M, Moser, M. T, Mostafavi, S, Nicolae, D. L, Pritchard, J, Qi, L, Ramsey, K, Rivas, M. A, Robles, B. E, Rohrer, D. C, Salvatore, M, Sammeth, M, Seleski, J, Shad, S, Siminoff, L. A, Stephens, M, Struewing, J, Sullivan, T, Sullivan, S, Syron, J, Tabor, D, Taherian, M, Tejada, J, Temple, G. F, Thomas, J. A, Thomson, A. W, Tidwell, D, Traino, H. M, Tu, Z, Valley, D. R, Volpi, S, Walters, G. D, Ward, L. D, Wen, X, Winckler, W, Wu, S, Zhu, J, Abdallah, A, Addington, A, Anderson, J. M, Bender, P. K, Cosentino, M, Diaz-Mayoral, N, Engel, T, Garci, F, Green, A, Hammond, T, Jaffe, K, Keen, J, Kennedy, M, Kigonya, P, Lander, B, Nampally, S, Ny, C, Robb, J, Santhanum, V, Sharopova, N, Singh, S, Soria, C, Sturcke, A, Sukari, S, Thomson, E. J, Tomaszewski, M, Trowbridge, C, Udoye, F, Vanscoy, D, Vatanian, N, Wilder, E. L, & Williams, P. (2015) *Biopreservation and Biobanking* **13**, 311–317.

[63] Aguet, F, Barbeira, A. N, Bonazzola, R, Brown, A, Castel, S. E, Jo, B, Kasela, S, Kim-Hellmuth, S, Liang, Y, Oliva, M, Flynn, E. D, Parsana, P, Fresard, L, Gamazon, E. R, Hamel, A. R, He, Y, Hormozdiari, F, Mohammadi, P, Muñoz-Aguirre, M, Park, Y. S, Saha, A, Segrè, A. V, Strober, B. J, Wen, X, Wucher, V, Ardlie, K. G, Battle, A, Brown, C. D, Cox, N, Das, S, Dermitzakis, E. T, Engelhardt, B. E, Garrido-Martín, D, Gay, N. R, Getz, G. A, Guigó, R, Handsaker, R. E, Hoffman, P. J, Im, H. K, Kashin, S, Kwong, A, Lappalainen, T, Li, X, MacArthur, D. G, Montgomery, S. B, Rouhana, J. M, Stephens, M, Stranger, B. E, Todres, E, Viñuela, A, Wang, G, Zou, Y, Anand, S, Gabriel, S, Graubert, A, Hadley, K, Huang, K. H, Meier, S. R, Nedzel, J. L, Nguyen, D. T, Balliu, B, Conrad, D. F, Cotter, D. J, deGoede, O. M, Einson, J, Eskin, E, Eulalio, T. Y, Ferraro, N. M, Gloudemans, M. J, Hou, L, Kellis, M, Li, X, Mangul, S, Nachun, D. C, Nobel, A. B, Park, Y, Rao, A. S, Reverter, F, Sabatti, C, Skol, A. D, Teran, N. A, Wright, F, Ferreira, P. G, Li, G, Melé, M, Yeger-Lotem, E, Barcus, M. E, Bradbury, D, Krubit, T, McLean, J. A, Qi, L, Robinson, K, Roche, N. V, Smith, A. M, Sobin, L, Tabor, D. E, Undale, A, Bridge, J, Brigham, L. E, Foster, B. A, Gillard, B. M, Hasz, R, Hunter, M, Johns, C, Johnson, M, Karasik, E, Kopen, G, Leinweber, W. F, McDonald, A, Moser, M. T, Myer, K, Ramsey, K. D, Roe, B, Shad, S, Thomas, J. A, Walters, G, Washington, M, Wheeler, J, Jewell, S. D, Rohrer, D. C, Valley, D. R, Davis, D. A, Mash, D. C, Branton, P. A, Sobin, L, Barker, L. K, Gardiner, H. M, Mosavel, M, Siminoff, L. A, Flicek, P, Haeussler, M, Juettemann, T, Kent, W. J, Lee, C. M, Powell, C. C, Rosenbloom, K. R, Ruffier, M, Sheppard, D, Taylor, K, Trevanion, S. J, Zerbino, D. R, Abell, N. S, Akey, J, Chen, L, Demanelis, K, Doherty, J. A, Feinberg, A. P, Hansen, K. D, Hickey, P. F, Hou, L, Jasmine, F, Jiang, L, Kaul, R, Kibriya, M. G, Li, J. B, Li, Q, Lin, S, Linder, S. E, Pierce, B. L, Rizzardi,

L. F, Smith, K. S, Snyder, M, Stamatoyannopoulos, J, Tang, H, Wang, M, Branton, P. A, Carithers, L. J, Guan, P, Koester, S. E, Little, A. R, Moore, H. M, Nierras, C. R, Rao, A. K, Vaught, J. B, & Volpi, S. (2020) *Science* **369**, 1318–1330.

[64] Mostafavi, H, Spence, J. P, Naqvi, S, & Pritchard, J. K. (2023) *Nature Genetics 2023 55:11* **55**, 1866–1875.

[65] Geiger, T, Cox, J, & Mann, M. (2010) *PLOS Genetics* **6**, e1001090.

[66] Dephoure, N, Hwang, S, O'Sullivan, C, Dodgson, S. E, Gygi, S. P, Amon, A, & Torres, E. M. (2014) *eLife* **3**.

[67] Wang, D, Eraslan, B, Wieland, T, Hallström, B, Hopf, T, Zolg, D. P, Zecha, J, Asplund, A, hua Li, L, Meng, C, Frejno, M, Schmidt, T, Schnatbaum, K, Wilhelm, M, Ponten, F, Uhlen, M, Gagneur, J, Hahne, H, & Kuster, B. (2019) *Molecular Systems Biology* **15**, e8503.

[68] Franks, A, Airoldi, E, & Slavov, N. (2017) *PLoS Computational Biology* **13**, e1005535.

[69] Brion, C, Caradec, C, Pflieger, D, Friedrich, A, & Schacherer, J. (2020) *Molecular biology and evolution* **37**, 2520–2530.

[70] Stern, D. L & Orgogozo, V. (2008) The loci of evolution: How predictable is genetic evolution?

[71] Villar, D, Flicek, P, & Odom, D. T. (2014) *Nature Reviews Genetics 2014 15:4* **15**, 221–233.

[72] Osterwalder, M, Barozzi, I, Tissiéres, V, Fukuda-Yuzawa, Y, Mannion, B. J, Afzal, S. Y, Lee, E. A, Zhu, Y, Plajzer-Frick, I, Pickle, C. S, Kato, M, Garvin, T. H, Pham, Q. T, Harrington, A. N, Akiyama, J. A, Afzal, V, Lopez-Rios, J, Dickel, D. E, Visel, A, & Pennacchio, L. A. (2018) *Nature* **554**, 239–243.

[73] Kaplow, I. M, Schäffer, D. E, Wirthlin, M. E, Lawler, A. J, Brown, A. R, Kleyman, M, & Pfenning, A. R. (2022) *BMC Genomics 2022 23:1* **23**, 1–23.

[74] Kaplow, I. M, Sestili, H. H, Prasad, K, Brown, A. R, Foley, K, Pfenning, A. R, Andrews, G, Armstrong, J. C, Bianchi, M, Birren, B. W, Bredemeyer, K. R, Breit, A. M, Christmas, M. J, Clawson, H, Damas, J, Palma, F. D, Diekhans, M, Dong, M. X, Eizirik, E, Fan, K, Fanter, C, Foley, N. M, Forsberg-Nilsson, K, Garcia, C. J, Gatesy, J, Gazal, S, Genereux, D. P, Goodman, L, Grimshaw, J, Halsey, M. K, Harris, A. J, Hickey, G, Hiller, M, Hindle, A. G, Hubley, R. M, Hughes, G. M, Johnson, J, Juan, D, Kaplow, I. M, Karlsson, E. K, Keough, K. C, Kirilenko, B, Koepfli, K. P, Korstian, J. M, Kowalczyk, A, Kozyrev, S. V, Lawler, A. J, Lawless, C, Lehmann, T, Levesque, D. L, Lewin, H. A, Li, X, Lind, A, Lindblad-Toh, K, Mackay-Smith, A, Marinescu, V. D, Marques-Bonet, T, Mason, V. C, Meadows, J. R, Meyer, W. K, Moore, J. E, Moreira, L. R, Moreno-Santillan, D. D, Morrill, K. M, Muntané, G, Murphy, W. J, Navarro, A, Nweeia, M, Ortmann, S, Osmanski, A, Paten, B, Paulat, N. S, Pfenning, A. R, Phan, B. D. N, Pollard, K. S, Pratt, H. E, Ray, D. A, Reilly, S. K, Rosen, J. R, Ruf, I, Ryan, L, Ryder, O. A, Sabeti, P. C, Schäffer, D. E, Serres, A, Shapiro, B, Smit, A. F, Springer, M, Srinivasan, C, Steiner, C, Storer, J. M, Sullivan, K. A, Sullivan, P. F, Sundström, E, Supple, M. A, Swofford, R, Talbot, J. E, Teeling, E,

Turner-Maier, J, Valenzuela, A, Wagner, F, Wallerman, O, Wang, C, Wang, J, Weng, Z, Wilder, A. P, Wirthlin, M. E, Xue, J. R, & Zhang, X. (2023) *Science* **380**.

[75] Hansen, T. F, Pienaar, J, & Orzack, S. H. (2008) *Evolution* **62**, 1965–1977.

[76] Aubé, S, Nielly-Thibault, L, & Landry, C. R. (2023) *PLOS Genetics* **19**, e1010756.

[77] Halligan, D. L & Keightley, P. D. (2009) *Annual Review of Ecology, Evolution, and Systematics* **40**, 151–172.

[78] Landry, C. R, Lemos, B, Rifkin, S. A, Dickinson, W. J, & Hartl, D. L. (2007) *Science* **317**, 118–121.

[79] Hodgins-Davis, A, Duveau, F, Walker, E. A, & Wittkopp, P. J. (2019) *Proceedings of the National Academy of Sciences of the United States of America* **116**, 21085–21093.

[80] Frankel, N, Davis, G. K, Vargas, D, Wang, S, Payre, F, & Stern, D. L. (2010) *Nature 2010 466:7305* **466**, 490–493.

[81] Payne, J. L & Wagner, A. (2014) *Science* **343**.

[82] Berthelot, C, Villar, D, Horvath, J. E, Odom, D. T, & Flicek, P. (2017) *Nature Ecology and Evolution* **2**, 152–163.

[83] Tsai, A, Alves, M. R, & Crocker, J. (2019) *eLife* **8**.

[84] Wagner, A. (2008) *Proceedings of the Royal Society B: Biological Sciences* **275**.

[85] Raj, A & van Oudenaarden, A. (2008) *Cell* **135**, 216–226.

[86] Dunn, C. W, Zapata, F, Munro, C, Siebert, S, & Hejnol, A. (2018) *Proceedings of the National Academy of Sciences of the United States of America* **115**, E409–E417.

[87] Fukushima, K & Pollock, D. D. (2020) *Nature Communications* **11**, 1–14.

[88] Cope, A. L, O'Meara, B. C, & Gilchrist, M. A. (2020) *BMC Genomics* **21**, 370.

[89] Petit, A. J, Guez, J, & Rouzic, A. L. (2023) *Genetics* **224**, 65.

[90] Bedford, T & Hartl, D. L. (2009) *Proceedings of the National Academy of Sciences of the United States of America* **106**, 1133–8.

[91] Nourmohammad, A, Rambeau, J, Held, T, Kovacova, V, Berg, J, & Lässig, M. (2017) *Cell Reports* **20**, 1385–1395.

[92] Lynch, M. (1990) *American Naturalist* **136**, 727–741.

[93] Uyeda, J. C, Hansen, T. F, Arnold, S. J, & Pienaar, J. (2011) *Proceedings of the National Academy of Sciences of the United States of America* **108**, 15908–15913.

[94] Harmon, L. J, Pennell, M. W, Henao-Diaz, L. F, Rolland, J, Sipley, B. N, & Uyeda, J. C. (2021) Causes and consequences of apparent timescaling across all estimated evolutionary rates.

[95] Schraiber, J. G, Edge, M. D, & Pennell, M. (2024) *bioRxiv*.