Dissecting genomic determinants of positive selection with an evolution-guided regression model

Yi-Fei Huang[1,2]

[1]Department of Biology, Pennsylvania State University,
University Park, PA 16802, USA

[2]Huck Institutes of the Life Sciences, Pennsylvania State University,
University Park, PA 16802, USA

Abstract

In evolutionary genomics, it is fundamentally important to understand how characteristics of genomic sequences, such as the expression level of a gene, determine the rate of adaptive evolution. While numerous statistical methods, such as the McDonald-Kreitman test, are available to examine the association between genomic features and positive selection, we currently lack a statistical approach to disentangle the direct effects of genomic features from the indirect effects mediated by confounding factors. To address this problem, we present a novel statistical model, the MK regression, which augments the McDonald-Kreitman test with a generalized linear model. Analogous to the classic multiple regression model, the MK regression can analyze multiple genomic features simultaneously to distinguish between direct and indirect effects on positive selection. Using the MK regression, we identify numerous genomic features responsible for positive selection in chimpanzees, including local mutation rate, residue exposure level, gene expression level, tissue specificity, and metabolic genes. In particular, we show that highly expressed genes have a higher rate of adaptation than their weakly expressed counterparts, even though a higher expression level may impose stronger negative selection on protein sequences. Also, we observe that metabolic genes tend to have a higher rate of adaptation than their non-metabolic counterparts, possibly due to recent changes in diet in primate evolution. Overall, the MK regression is a powerful approach to elucidate the genomic basis of adaptation.

Significance Statement

The McDonald-Kreitman test is a powerful method to detect genomic features associated with positive selection, but it cannot disentangle the direct effects of genomic features from the indirect effects mediated by confounding factors. Here we introduce the MK regression that augments the McDonald-Kreitman test with a generalized linear model. Analogous to the classic multiple regression model, the MK regression enables quantification of the direct effects of genomic features by controlling for potential confounders. It is particularly useful to analyze data with multiple correlated features and species with low levels of polymorphism. Using the MK regression, we identify numerous genomic features responsible for positive selection in chimpanzees, which provides novel insights into the adaptive evolution of our closest relatives.

Introduction

Understanding the genetic basis of positive selection is a fundamental problem in evolutionary biology. Numerous statistical approaches have been developed to detect loci under positive selection. A popular framework is codon substitution models that seek to infer positively selected genes solely from interspecies sequence divergence [1–3]. By contrasting the rate of nonsynonymous substitutions (dN) against the rate of synonymous substitutions (dS), codon substitution models can identify positively selected genes with a dN/dS ratio greater than 1. However, because negative (purifying) selection can dramatically reduce dN/dS ratios, codon substitution models may be underpowered to detect genes that experienced both positive selection and strong negative selection [4].

Unlike codon substitution models, the McDonald-Kreitman (MK) test utilizes both interspecies divergence and intraspecies polymorphism to elucidate positive selection in a species of interest [5–7]. By contrasting the levels of divergence and polymorphism at functional sites and putatively neutral sites, the MK test seeks to identify positively selected genes that show an excess of interspecies divergence at functional sites. Because strong negative selection has similar effects on divergence and polymorphism data, the MK test is intrinsically robust to the presence of strong negative selection. Also, several recent studies have extended the MK test to account for the effects of weak negative selection on intraspecies polymorphism [8–12], ensuring that the inference of positive selection is not biased by the presence of weak selection. Thus, the MK test and its extensions are powerful methods to disentangle positive selection from ubiquitous negative selection.

Because MK-based methods only use divergence and polymorphism data from one species, they may be underpowered to pinpoint individual genes under positive selection. To boost statistical power, MK-based methods often are applied to a collection of genes or nucleotide sites with similar genomic features. Using this pooling strategy, previous studies have identified numerous genomic features associated with positive selection, including local mutation rate [13–15], local recombination rate [13, 14, 16], gene expression specificity [17], residue exposure to solvent [18], protein disorder [19], virus-host interaction [12, 20], protein-protein interaction (PPI) degree [21], X linkage [22–24], and sex-biased expression [23–25].

While existing MK-based methods can identify genomic features *associated with* the signatures of positive selection, they may not be able to distinguish genomic features *responsible for* positive selection from

spurious features *without direct effects* on adaptation [17, 18]. For instance, MK-based methods often are applied to one genomic feature at a time. If a genomic feature responsible for positive selection is strongly correlated with a second feature without direct effect, MK-based methods may report a spurious association between the second feature and adaptive evolution.

To address this problem, two strategies have been previously developed to estimate the direct effect of a genomic feature on positive selection by controlling for the effects of potential confounding factors. If we are interested in estimating the effect of a gene-level feature, such as tissue specificity, we may estimate the rate of adaption at the gene level and then fit a standard linear regression model, in which we treat the feature of interest and potential confounders as covariates and treat the gene-level rate of adaptation as a response variable [14, 17, 18, 21]. The regression coefficient associated with the feature of interest can be interpreted as its direct effect on positive selection after controlling for confounders. Although this strategy is powerful and elegant, it cannot be applied to species with low levels of polymorphism, such as primates, due to the challenge of estimating the rate of adaptation at the gene level. Alternatively, we may first stratify genes into a "treatment" group and a "control" group based on the genomic feature of interest. Then, we may use statistical matching algorithms to match each gene from the "treatment" group with a gene of similar characteristics from the "control" group. A significant difference in the rate of adaptation between the two groups of matched genes indicates that the feature of interest has a direct effect on positive selection. Although this method has been successfully used in previous studies [20, 24, 26], it is difficult to match genes when there are a large number of confounders to control for. Therefore, we currently lack a general and powerful statistical framework to estimate the direct effects of genomic features on positive selection by adjusting for a large number of potential confounders.

In the current study, we present a novel statistical method, the MK regression, to estimate the direct effects of genomic features on the rate of adaptive evolution. The MK regression is a hybrid of the MK test and the generalized linear regression. Unlike standard linear regression models and statistical matching algorithms, the MK regression can control for a large number of confounding factors and is applicable to species with a low level of polymorphism. To the best of our knowledge, the MK regression is the first evolutionary model tailored to characterize the direct effects of genomic features on adaptive evolution. Using synthetic data, we show that the MK regression can unbiasedly estimate the direct effect of a genomic feature even when it is strongly correlated with a confounding factor. Using the MK regression, we show

that local mutation rate, residue exposure level, gene expression level, tissue specificity, and metabolic genes are key determinants of positive selection in the chimpanzee lineage. Taken together, the MK regression is a valuable addition to evolutionary biologists' arsenal for investigating the genetic basis of adaptation.

Results

The MK regression is a generalized linear model tailored to estimate the effects of genomic features on positive selection

The key idea behind the MK regression is to model the site-wise rate of adaptive evolution as a linear combination of local genomic features (Fig. 1). We use ω_α, the relative rate of adaptive substitutions at a functional site with respect to the average substitution rate at neutral sites, as a measure of the rate of adaptation [27, 28]. Unlike previous MK-based models that treat ω_α as a gene-level measure, we treat ω_α as a site-level measure and assume that it can be predicted from local genomic features. To integrate the effects of multiple features on the rate of adaptive evolution, we assume that ω_α, in a site-wise manner, is a linear combination of local genomic features, such as local mutation rate, local recombination rate, and gene expression level. For each genomic feature, the MK regression seeks to estimate a regression coefficient indicating its direct effect on the rate of adaption after controlling for the effects of other features.

Specifically, the MK regression consists of two components: a generalized linear model and a MK-based likelihood function (Fig. 1). First, we assume that ω_α, in a site-wise manner, is a linear combination of local genomic features followed by an exponential transformation,

$$\omega_\alpha = \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i + \cdots + \beta_M X_M).$$

In this equation, X_i is the i-th feature at a functional site; β_0 is an intercept indicating the baseline rate of adaptive evolution when all genomic features are equal to 0; β_i is a regression coefficient indicating the i-th feature's effect on the rate of adaptation; M is the total number of genomic features; $\exp$ is an exponential inverse link function which ensures ω_α is positive. If β_i is statistically different from 0, we consider that feature i may have a direct effect on adaptation after adjusting for the other features. Similarly, to accommodate the effects of local genomic features on polymorphism data, we assume that the probability of observing intraspecies polymorphism at a functional site, P_{func}, is another linear combination of local

genomic features followed by a logistic transformation (Fig. 1).

Second, in the component of MK-based likelihood function (Fig. 1), we combine ω_α and P_{func} at every functional site with two neutral parameters to calculate the probability of observed polymorphism and divergence data at both functional and neutral sites, which allows for a maximum likelihood estimation of model parameters. Finally, we use the Wald test to examine whether the estimated regression coefficient, $\hat{\beta}_i$, is significantly different from 0 for each feature i.

Joint analysis of multiple features distinguishes direct effects from spurious associations

We conducted two simulation experiments to assess the MK regression's validity and its power to distinguish genomic determinants *responsible for* positive selection from spurious features *indirectly associated with* selection signals. The simulation experiments consisted of two steps. First, we randomly sampled genomic features from a bivariate normal distribution at each functional site. Second, we generated synthetic polymorphism and divergence data at both functional and neutral sites based on the MK regression model.

In the first simulation experiment, we assumed that there were two genomic features of interest. The first genomic feature was responsible for positive selection, and its correlation coefficient, β_1, was equal to 1. On the other hand, the second feature had no direct effect on selection. Thus, its regression coefficient, β_2, was equal to 0 by definition. The other parameters required for the simulation experiment were chosen to ensure that genome-wide levels of polymorphism and divergence are comparable between synthetic data and empirical data from chimpanzees (see details in the Materials and Methods section). To systematically assess the MK regression's performance with respect to various degrees of correlation between genomic features, we generated four sets of synthetic data with different correlation coefficients between features (0.0, 0.2, 0.4, and 0.6). In each synthetic dataset, we generated 10 independent replicates each of which consisted of 10 Mb functional sites and 10 Mb neutral sites.

We applied two different versions of the MK regression to the synthetic data. The first version was the simple MK regression that analyzed one genomic feature at a time, which was designed to mimic previous MK-based methods. The second one was the multiple MK regression that analyzed two features simultaneously. As shown in Fig. 2A & B, both the simple MK regression and the multiple MK regression produced unbiased estimates of regression coefficients when there was no correlation between features. However, the simple MK regression frequently estimated that $\hat{\beta}_2$ was positive when the two features were correlated with

each other, whereas the true value of β_2 was equal to 0. On the other hand, the multiple MK regression always produced unbiased estimates of regression coefficients regardless of the degree of correlation between features.

In the second simulation experiment, we evaluated the extent to which the correlation between two causal features complicates the estimation of their direct effects. We set the regression coefficients of the two features to $\beta_1 = 1$ and $\beta_2 = -0.2$, respectively. Then, we followed the same procedure described in the first simulation experiment to generate synthetic data. As shown in Fig. 2C & D, the multiple MK regression accurately estimated regression coefficients without any noticeable bias, whereas the simple MK regression produced biased results when the two features were correlated with each other. Importantly, when the correlation was strong, the simple MK regression estimated that $\hat{\beta}_2$ was positive while the true value of β_2 was equal to -0.2. In summary, it is critical to jointly analyze multiple genomic features for an unbiased estimation of their direct effects on adaptive evolution.

The multiple MK regression elucidates genomic determinants of positive selection in chimpanzees

We investigated positive selection in chimpanzee autosomal genes using the MK regression. Because gene annotations were of high quality in the human genome, we converted 4-fold degenerate (4D) and 0-fold degenerate (0D) sites annotated in dbNSFP [29, 30] from the human genome to the chimpanzee genome. Because all point mutations at 4D sites are synonymous, we assume that they are putatively neutral. On the other hand, because all point mutations at 0D sites are nonsynonymous, we assume that they are potentially functional. We obtained a genome-wide map of single nucleotide polymorphisms (SNPs) in 18 central chimpanzee (*Pan troglodytes troglodytes*) individuals [31], and inferred ancestral alleles using a reconstructed chimpanzee ancestral genome [32, 33]. To mitigate the impact of weak negative selection on the inference of positive selection, we filtered out SNPs with a derived allele frequency lower than 50% for downstream analysis. In addition, we reconstructed fixed substitutions at 4D and 0D sites in the chimpanzee lineage by comparing the reconstructed ancestral genome with the chimpanzee reference genome. We estimated that the proportion of adaptive nonsynonymous substitutions (α) was equal to 15.3% in chimpanzee autosomal genes, which is similar to the estimate in a previous study [34].

We collected six genomic features in chimpanzee autosomal genes, including local mutation rate, lo-

cal recombination rate, residue exposure level, gene expression level, tissue specificity, and the number of unique protein-protein-interaction partners per gene (PPI degree). Specifically, we obtained a map of local recombination rates from a previous study [35] and constructed a map of local mutation rates using putatively neutral substitutions in the chimpanzee lineage. We obtained tissue-based gene expression data from the Human Protein Atlas [36], and utilized the expression level averaged across all tissues and a summary statistic, tau [37], as measures of gene expression level and tissue specificity, respectively. We also obtained predicted levels of residue exposure to solvent and experimentally determined PPI degrees from previous studies [38, 39].

We first employed the simple MK regression to analyze the effect of one genomic feature at a time, with no attempt to distinguish direct effects from spurious associations. Because the MK regression used a logarithmic link function for ω_α, we explore if a logarithmic transformation of genomic features can improve model fitting. We found that the logarithmic transformation improved the fitting of the simple MK regression for all features but tissue specificity (Table S1). Therefore, we applied the logarithmic transformation to all features except tissue specificity throughout this study. In the simple MK regression, the regression coefficients of local mutation rate, residue exposure level, and PPI degree were significantly higher than 0, whereas the regression coefficient of gene expression level was significantly lower than 0 (Fig. 3A & Table S2). On the other hand, local recombination rate and tissue specificity were not significantly associated with rate of adaptive evolution in the simple MK regression (Fig. 3A & Table S2).

As discussed in the simulation experiments, the simple MK regression may produce biased estimates of regression coefficients if genomic features are correlated with each other. To test if this was the case in our chimpanzee data, we used the multiple MK regression to analyze the effects of the six genomic features simultaneously. Surprisingly, while local mutation rate, local recombination rate, and residue exposure level showed similar effects in the multiple MK regression, the regression coefficients of the other features were different between the multiple MK regression and the simple MK regression (Fig. 3B & Table S3). Specifically, the regression coefficient of PPI degree was not significant in the multiple MK regression ($\hat{\beta}$ = 0.002; *p*-value = 0.975), whereas the same coefficient was significant in the simple MK regression ($\hat{\beta}$ = 0.460; *p*-value = $3.178 \cdot 10^{-4}$). The coefficient of gene expression level was significantly higher than 0 in the multiple MK regression ($\hat{\beta}$ = 0.347; *p*-value = $6.674 \cdot 10^{-12}$), whereas the same coefficient was negative in the simple MK regression ($\hat{\beta}$ = -0.310; *p*-value = $3.534 \cdot 10^{-10}$). Also, the regression coefficient of tissue

specificity was significantly higher than 0 in the multiple MK regression ($\hat{\beta}$ = 1.168; p-value = $6.394 \cdot 10^{-31}$) but not in the simple MK regression ($\hat{\beta}$ = 0.157; p-value = 0.622).

To investigate whether correlations between genomic features could explain the differences in estimated coefficients between the multiple and the simple MK regression, we calculated the Kendall rank correlation coefficient for all pairs of genomic features (Fig. 3C & Table S4). We found that local mutation rate, local recombination rate, and residue exposure level were weakly correlated with other features, which may explain why the regression coefficients of these features were consistent between the multiple and the simple MK regression. In contrast, gene expression level and tissue specificity showed a strong negative correlation, which may cause spurious associations in the simple MK regression. We also found that PPI degree was correlated with gene expression level and tissue specificity, although the correlations were to a lesser extent compared with the correlation between gene expression level and tissue specificity. Thus, the observed association of PPI degree with positive selection in the simple MK regression could be due to an indirect effect mediated by gene expression level and/or tissue specificity.

Statistical matching analysis confirms genomic determinants identified by the multiple MK regression

We used statistical matching algorithms to corroborate the results of the multiple MK regression. First, we verified whether the association between PPI degree and the rate of adaptive evolution was mediated by gene expression level and/or tissue specificity. We stratified protein-coding genes into two groups with different PPI degrees, in which 1,556 genes with at least 10 protein interaction partners were assigned to the high PPI-degree group whereas 8,471 genes with no more than one interaction partner were assigned to the low PPI-degree group. Without controlling for gene expression level and tissue specificity, the high PPI-degree group had a higher ω_α than the low PPI-degree group (Fig. S1A), but the difference in ω_α was not significant (p-value = 0.146; two-tailed permutation test), possibly due to a reduction of sample size in the stratified analysis. Then, using the default propensity score matching algorithm in MatchIt [40], we matched each gene from the high PPI-degree group with a gene of similar expression level and tissue specificity from the low PPI-degree group. In the matched data, ω_α was not different between the high-PPI and low-PPI groups (Fig. S1B). Thus, PPI degree is unlikely to be a genomic determinant of positive selection in chimpanzees.

We also verified the effect of tissue specificity on the rate of adaptation after adjusting for gene expres-

sion level. Due to the strong negative correlation between expression level and tissue specificity (Fig. S2), MatchIt returned few matched genes when we attempted to control for expression level. Therefore, we implemented a different matching approach. By closely examining the relationship between expression level and tissue specificity, we found that the variation of tissue specificity was high among highly expressed genes (Fig. S2). Thus, we stratified 737 highly expressed genes (gene expression level > 30) into two equal-sized groups based on tissue specificity. As shown in Fig. 4A, highly expressed genes with high tissue specificity had a significantly higher ω_α than their counterparts with low tissue specificity (p-value = 0.001; two-tailed permutation test). Therefore, our gene matching analysis confirms the positive effect of tissue specificity on the rate of adaptation in the multiple MK regression.

Similarly, we observed that the variation of expression level was high among genes with high tissue specificity (Fig. S2). To verify the positive effect of gene expression level after adjusting for tissue specificity, we stratified 993 tissue-specific genes (tau > 0.85) into two approximately equal-sized groups based on expression level. The first group consisted of 495 tissue-specific genes with high expression level, whereas the second group consisted of 498 tissue-specific genes with low expression level. As shown in Fig. 4B, tissue-specific genes with high expression level had a significantly higher ω_α than their lowly expressed counterparts (p-value = 0.002; two-tailed permutation test), which confirms the positive effect of gene expression level on the rate of adaptive evolution in the multiple MK regression.

Also, we used phyloP scores [41, 42] to examine the effects of gene expression level and tissue specificity on the rate of protein evolution. After controlling for gene expression level, phyloP scores increased with decreasing tissue specificity (Fig. 4C), which is in line with the observation that housekeeping genes tend to evolve at a lower substitution rate than tissue-specific genes [43, 44]. On the other hand, after controlling for tissue specificity, phyloP scores increased with increasing expression level (Fig. 4D), which is in line with stronger purifying selection on highly expressed genes [45]. Taken together, it seems that highly expressed genes may be subject to more frequent positive selection than their lowly expressed counterparts, although a higher expression level may also impose stronger purifying selection and tends to reduce the overall rate of protein evolution.

Metabolic genes are under frequent positive selection in chimpanzees

To explore whether the positive effect of gene expression level on adaptive evolution can be explained by the functions of highly expressed genes, we examined the enrichment of Reactome pathways [46] and tissue types [36] in the aforementioned 495 tissue-specific genes with high expression level, using the 498 tissue-specific genes with low expression level as a background set. In line with frequent positive selection on immune system [47–50] and male-reproductive system [23–25, 48, 51], we observed that genes associated with immune system and genes with enriched expression in epididymis had a 2- to 4-fold enrichment in the 495 tissue-specific genes with high expression level (Fig. 5A & B; false-discovery rate < 0.01). Thus, frequent positive selection on immune system and male-reproductive system might partially explain the positive effect of gene expression level on the rate of adaptive evolution.

Surprisingly, genes associated with the metabolism of proteins and lipids, and genes with enriched expression in intestine, liver, and pancreas, showed an even stronger enrichment in the 495 tissue-specific genes with high expression level (Fig. 5A & B; false-discovery rate < 0.01). Based on these results, we hypothesized that metabolic genes may be subject to more frequent positive selection than non-metabolic genes in chimpanzees. To test this hypothesis, we constructed a new genomic feature indicating whether each 0D site was located in one of the 2,220 metabolic genes from MSigDB [52, 53]. Then, we used the multiple MK regression to simultaneously estimate the effects of the new feature and the six original features. As shown in Fig. 5C & Table S5, the regression coefficient of metabolic genes was significantly higher than 0 in the multiple MK regression ($\hat{\beta}$ = 0.537; *p*-value = $1.075 \cdot 10^{-4}$), suggesting that metabolic genes indeed had a higher rate of adaptation than their non-metabolic counterparts. Interestingly, the effect of metabolic genes was not significant in the simple MK regression ($\hat{\beta}$ = 0.276; *p*-value = 0.516). Therefore, controlling for potential confounders is critical for revealing elevated positive selection in metabolic genes. Metabolic genes may partially explain the positive effect of gene expression level on adaptive evolution, because the regression coefficient of gene expression level reduced moderately from 0.347 to 0.314 after adding metabolic genes as a new feature in the multiple MK regression (Table S3 & S5).

Discussion

In this work, we have introduced the MK regression, the first evolutionary model for jointly estimating the effects of multiple, potentially correlated genomic features on the rate of adaptive substitutions. Based on similar ideas, we have previously developed statistical approaches to infer negative selection on genetic variants [54–56]. Thus, unifying generalized linear models and evolutionary models may be a powerful strategy to address a variety of statistical problems in evolutionary biology.

As shown in the simulation experiments (Fig. 2), when two genomic features are correlated with each other, even at a moderate level, the simple MK regression cannot accurately estimate the direct effects of genomic features because it cannot control for confounding factors. On the other hand, the multiple MK regression can unbiasedly estimate the direct effects of genomic features if potential confounding factors are included in the same analysis. Because we are almost always interested in the direct effects of genomic features, the multiple MK regression may be superior to the simple MK regression and other MK-based methods that can only analyze one feature at a time.

Regression coefficients in the multiple MK regression might be interpreted as causal effects of genomic features on the rate of adaptation. However, similar to other linear regression models [57], the causal interpretation of the multiple MK regression relies on two implicit assumptions. First, genomic features of interest and all potential confounding factors should be included in the same MK regression analysis. Second, there should be no reverse causality, *i.e.*, natural selection should not cause changes in genomic features. As discussed in the literature of causal inference [57], these assumptions cannot be verified using observational data alone and, thus, have to be justified by domain knowledge in a case-by-case basis.

Using the multiple MK regression, we have identified numerous genomic features responsible for adaptive evolution in the chimpanzee lineage (Fig. 3B). First, in line with previous studies [13–15], we have shown that the rate of adaptation increases with increasing mutation rate. Because mutations are the ultimate source of genetic variation, a higher mutation rate will increase the genetic variation for positive selection to act on.

Second, previous studies have shown that local recombination rate is positively correlated with the rate of adaptation in *Drosophila* [13, 14, 16], probably due to a reduced effect of Hill-Robertson interference in recombination hotspots. However, we have not observed the same pattern in chimpanzees, which could be

explained by a reduced impact of linked selection in species with a small census population size, such as chimpanzees [58]. Alternatively, our analysis may have limited power to detect a weak association between recombination rate and positive selection due to the lower level of polymorphism and the smaller proportion of adaptive substitutions in chimpanzees compared with *Drosophila* [14].

Third, in agreement with a previous study [18], we have shown that the rate of adaptive evolution increases with the increasing level of residue exposure to solvent. On the other hand, it is well known that the site-wise rate of protein evolution is positively correlated with the level of residue exposure, possibly due to relaxed negative selection on exposed residues [59–62]. Taken together, exposed residues on protein surfaces may be subject to both weaker negative selection and more frequent positive selection than their buried counterparts. From a biophysical perspective, missense mutations on protein surfaces are less likely to disrupt protein stabilities than mutations in hydrophobic cores [60, 63–65]. Thus, missense mutations on protein surfaces may be under more frequent positive selection because they are less likely to perturb protein folding. Alternatively, protein surfaces may have a higher rate of adaptive evolution because they may play an important role in host-pathogen interactions [18].

Fourth, we have shown that the tissue specificity of a gene has a positive effect on the rate of adaptation after controlling for potential confounders, such gene expression level. Thus, tissue-specific genes are more likely to be under positive selection than housekeeping genes. Because nonsynonymous mutations in housekeeping genes have a higher chance to disrupt multiple phenotypes, our findings support that pleiotropic effect is a key determinant of adaptive evolution [17].

Fifth, we have shown that the rate of adaptive evolution increases with increasing gene expression level after controlling for potential confounders, such as tissue specificity. Controlling for potential confounders is critical for determining the positive effect of expression level on adaptive evolution, because gene expression level is negatively associated with the rate of adaptation in the simple MK regression.

Last but not least, we have shown that highly expressed genes are more likely to be associated with metabolic pathways and digestive organs than their lowly expressed counterparts (Fig. 5A & B), which implies that frequent positive selection in metabolic genes may partially explain the positive effect of gene expression level on the rate of adaptation. In agreement with this hypothesis, we have shown that metabolic genes have a higher rate of adaptation than their non-metabolic counterparts after controlling for potential confounders (Fig. 5C). Similarly, a recent study has reported that metabolic pathways may be subject to

more frequent positive selection than non-metabolic pathways in multiple inner branches of the primate phylogeny [66]. Taken together, genes in metabolic pathways may be subject to frequent positive selection in multiple primate species, possibly due to recent changes in diet in primate evolution [66–69].

Frequent positive selection in metabolic genes has not been widely reported in primates, except in the current study and in [66]. We hypothesize that the discrepancy could be explained by the unique design of the MK regression and the method in [66]. First, previous studies focused on identifying individual genes with significant signals of positive selection. If metabolic pathways are under polygenic selection, the signal of selection in a single gene may be too weak to reach genome-wide significance [70, 71]. In contrast, the MK regression and the method in [66] have pooled data across a large number of metabolic genes, which may significantly increase the statistical power to detect diffused signals of polygenic selection [71]. Second, if metabolic genes are under lineage-specific adaptation in primates, positive selection may only be detected by statistical approaches tailored for a single branch of the primate phylogeny, such as the MK regression and the branch-site codon substitution model [66]. Third, many previous methods may not be able to control for the effects of potential confounders. In the current study, we have shown that the effect of metabolic genes is manifested in the multiple MK regression but not in the simple MK regression, which highlights the importance of controlling for potential confounders. In future studies, it is tempting to test these hypotheses for a comprehensive understanding of adaptive evolution in metabolic genes.

Similar to a previous study [21], we have found that the rate of adaptation increases with increasing PPI degree in the simple MK regression (Fig. 3A). However, the same pattern cannot be replicated in the multiple MK regression (Fig. 3B). Also, after controlling for gene expression level and tissue specificity using propensity score matching, we have found no difference in the rate of adaptation between genes with high PPI degree and genes with low PPI degree (Fig. S1B). Thus, PPI degree is unlikely to be a key determinant of positive selection in chimpanzees.

While the MK regression is a powerful framework to estimate the effects of multiple genomic features simultaneously, it has a few limitations that are worth of future exploration. The MK regression does not explicitly model the effects of weak selection on polymorphism data. To mitigate this problem, we have used a simple strategy to filter out low frequency SNPs. This strategy may not be optimal, because a large proportion of SNPs cannot be used in the MK regression despite the fact that they are potentially informative of positive selection. In future studies, it is tempting to extend the MK regression by explicitly modeling

the effects of weak selection on polymorphism data [8–12]. Another direction for future improvement is to relax the strong assumption of a linear relationship between genomic features and the rate of adaptation. For instance, we may replace the generalized linear model in the MK regression by a generalized additive model [72], which can accommodate more complicated relationships between features and selection while maintaining the interpretability of the MK regression. We expect that these extensions will empower the MK regression to explore the genomic basis of adaptive evolution in various species, such as humans and *Drosophila*.

Materials and Methods

Details of the MK regression

The MK regression consists of two components: a generalized linear model and a MK-based likelihood function. In the generalized linear model, we assume that ω_α^j, the relative rate of adaptive substitutions at functional site j, is a linear combination of genomic features followed by an exponential transformation,

$$\omega_\alpha^j = \exp(\beta_0 + \beta_1 X_1^j + \cdots + \beta_i X_i^j + \cdots + \beta_M X_M^j), \tag{1}$$

in which X_i^j is the i-th local genomic feature at site j, β_i is a regression coefficient indicating the effect of feature i on ω_α^j, β_0 in an intercept, and M is the total number of genomic features. Because genomic features may also influence the levels of polymorphism at functional sites, we model the probability of observing a SNP at functional site j, P_{func}^j, as another linear combination of genomic features followed by a logistic transformation,

$$\begin{aligned} P_{\text{func}}^j &= \text{logistic}(\gamma_0 + \gamma_1 X_1^j + \cdots + \gamma_i X_i^j + \cdots + \gamma_M X_M^j) \\ &= \frac{\exp(\gamma_0 + \gamma_1 X_1^j + \cdots + \gamma_i X_i^j + \cdots + \gamma_M X_M^j)}{1 + \exp(\gamma_0 + \gamma_1 X_1^j + \cdots + \gamma_i X_i^j + \cdots + \gamma_M X_M^j)}, \end{aligned} \tag{2}$$

in which γ_0 and γ_i are an intercept and a regression coefficient with respect to P_{func}^j, respectively. Similar to the regression coefficients in equation 1, γ_i represents the effect of feature i on the occurrence of polymorphism at functional site j. It is worth noting that we effectively assume an infinite-site model here [73], so no more than one SNP is allowed at a single site. Finally, we introduce two neutral parameters, D_{neut} and P_{neut}, which represent the expected number of substitutions and the probability of observing a SNP at a neutral site, respectively. These neutral parameters are shared by all neutral sites.

In the MK-based likelihood function, we specify the probability of polymorphism and divergence data at both neutral and functional sites given model parameters (β_0, β_i, γ_0, γ_i, D_{neut}, and P_{neut}). First, we denote Y^k_{neut} as a binary response variable indicating the presence/absence of a SNP at neutral site k and assume that it follows a Bernoulli distribution,

$$\mathbb{P}(Y^k_{\text{neut}}) = \begin{cases} P_{\text{neut}}, & \text{if } Y^k_{\text{neut}} = 1 \\ 1 - P_{\text{neut}}, & \text{otherwise.} \end{cases} \tag{3}$$

Similarly, we denote Y^j_{func} as a binary response variable indicating the presence/absence of a SNP at functional site j and assume that it follows a Bernoulli distribution,

$$\mathbb{P}(Y^j_{\text{func}}) = \begin{cases} P^j_{\text{func}}, & \text{if } Y^j_{\text{func}} = 1 \\ 1 - P^j_{\text{func}}, & \text{otherwise.} \end{cases} \tag{4}$$

Second, we employ the Jukes-Cantor substitution model [74] to describe the distribution of interspecies divergence at neutral sites. Denoting Z^k_{neut} as a binary response variable indicating if the reference genome and the ancestral genome have different nucleotides at neutral site k, the Jukes-Cantor model suggests that

$$\mathbb{P}(Z^k_{\text{neut}}) = \begin{cases} \frac{3}{4} - \frac{3}{4}\exp(-\frac{4}{3}D_{\text{neut}}), & \text{if } Z^k_{\text{neut}} = 1 \\ \frac{1}{4} + \frac{3}{4}\exp(-\frac{4}{3}D_{\text{neut}}), & \text{otherwise.} \end{cases} \tag{5}$$

Third, to model the effects of positive selection and neutral factors on interspecies divergence at functional sites, we assume that D^j_{func}, *i.e.*, the expected number of substitutions at functional site j, is equal to the sum of the number of adaptive substitutions and the number of neutral substitutions [75, 76],

$$D^j_{\text{func}} = \underbrace{\omega^j_\alpha D_{\text{neut}}}_{\substack{\text{Adaptive}\\ \text{substitutions}}} + \underbrace{\frac{P^j_{\text{func}}}{P_{\text{neut}}} D_{\text{neut}}}_{\substack{\text{Neutral}\\ \text{substitutions}}}, \tag{6}$$

in which $\frac{P^j_{\text{func}}}{P_{\text{neut}}}$ is equal to the relative rate of neutral evolution at functional site j compared with neutral sites. Given D^j_{func} at each functional site, we again employ the Jukes-Cantor substitution model to describe interspecies divergence at functional site j,

$$\mathbb{P}(Z^j_{\text{func}}) = \begin{cases} \frac{3}{4} - \frac{3}{4}\exp(-\frac{4}{3}D^j_{\text{func}}), & \text{if } Z^j_{\text{func}} = 1 \\ \frac{1}{4} + \frac{3}{4}\exp(-\frac{4}{3}D^j_{\text{func}}), & \text{otherwise,} \end{cases} \tag{7}$$

in which Z^j_{func} indicates if the reference genome and the ancestral genome have different nucleotides at functional site j. Finally, we assume that nucleotide sites evolve independently given genomic features and model parameters. Thus, we define the MK-based likelihood function of the whole dataset as

$$\prod_{j \in \text{all functional sites}} \mathbb{P}(Y^j_{\text{func}})\mathbb{P}(Z^j_{\text{func}}) \prod_{k \in \text{all neutral sites}} \mathbb{P}(Y^k_{\text{neut}})\mathbb{P}(Z^k_{\text{neut}}). \tag{8}$$

We estimate model parameters (β_0, $\underline{e}ta_i$, γ_0, γ_i, P_{neut}, and D_{neut}) by maximizing the logarithm of the MK-based likelihood function (equation 8). Also, we estimate the standard errors of parameters using the observed Fisher information matrix and compute the *p*-values of estimated parameters using the (two-tailed) Wald test.

Simulation experiments

In each simulation run, we first sampled genomic features (X^j_i) at 10 Mb functional sites from a bivariate normal distribution. We set the means and variances of the bivariate normal distribution to 0 and 1, respectively, and varied its correlation coefficient from 0 to 0.6 to examine the performance of the MK regression with respect to various degrees of correlation between features. Then, we computed ω^j_α, P^j_{func}, and D^j_{func} at each functional site j using equations 1, 2, and 6, respectively. Finally, we generated Y^k_{neut} and Z^k_{neut} at 10 Mb neutral sites and Y^j_{func} and Z^j_{func} at 10 Mb functional sites using equations 3, 4, 5, and 7.

We performed two simulation experiments each of which consisted of four sets of synthetic data. In the first experiment, we set $\beta_0 = -2$, $\beta_1 = 1$, $\beta_2 = 0$, $\gamma_0 = -8$, and sampled γ_1 and γ_2 from a normal distribution with a mean of 0 and a standard deviation of 0.5 in each simulation run. We generated four sets of synthetic data with different correlation coefficients (0, 0.2, 0.4, and 0.6) between features in the aforementioned bivariate normal distribution. In each dataset, we performed 10 independent simulation runs using the method described in the previous paragraph. In the second experiment, we generated synthetic data using the same procedure but replaced β_2 with -0.2.

Polymorphism and divergence data

Throughout this work, we focused on analyzing 0D and 4D sites within previously defined callable regions on autosomal chromosomes in the panTro4 reference genome [31]. We obtained whole genome sequencing (WGS) based genotypes of 18 central chimpanzee (*Pan troglodytes troglodytes*) individuals from [31]. We

filtered out all multiallelic sites and sites with missing genotypes. Then, we filtered out SNPs without a high-confidence ancestral allele (see below) and SNPs with a derived allele frequency below 0.5.

We obtained high-confidence chimpanzee ancestral alleles from Ensembl release 75 [32, 33]. Ensembl reconstructed chimpanzee ancestral alleles using a phylogenetic approach, and defined that an ancestral allele was of high confidence if it was identical to both the human reference allele and the reconstructed allele in the human-chimpanzee-macaque ancestor. To annotate interspecies divergence in the chimpanzee lineage, we compared the chimpanzee ancestral allele with the reference allele in panTro4 at each monoallelic site.

Estimation of α and ω_α

Based on the aforementioned polymorphism and divergence data in chimpanzee autosomal genes, we computed d_n, p_n, d_s, and p_s, which are the numbers of 0D sites with divergence, 0D sites with polymorphism, 4D sites with divergence, and 4D sites with polymorphism, respectively. We estimated the proportion of adaptive substitutions as [7, 77]

$$\alpha = 1 - \frac{d_s}{d_n}\frac{p_n}{p_s}. \tag{9}$$

Similarly, given the polymorphism and divergence data in a gene group of interest, we estimated the relative rate of adaptive substitutions with respect to neutral evolution as [75, 76]

$$\begin{aligned} \omega_\alpha &= \frac{d_n \cdot \alpha}{d_s} \\ &= \frac{d_n}{d_s} - \frac{p_n}{p_s}, \end{aligned} \tag{10}$$

in which we used 0D sites in the gene group of interest and 4D sites in all protein-coding genes as functional and neutral sites, respectively.

Genomic features

We obtained coordinates of 0D and 4D sites in the human genome from dbNSFP version 4.0 [29, 30], and converted the coordinates from the original hg19 assembly to the panTro4 assembly using liftOver [78]. We obtained predicted probabilities of residue exposure (PredRSAE) from SNVBox [38] and converted the data from hg19 to panTro4 with liftOver. We obtained consensus RNA expression levels across 62 tissues from the Human Protein Atlas version 19.3 [36]. For each protein-coding gene, we computed its (mean)

expression level,

$$\text{expression level} = \frac{\sum_1^K R_k}{K}, \tag{11}$$

in which R_k is the gene's consensus RNA expression level in tissue k and K is the total number of tissues. Also, we computed the tissue specificity of each gene,

$$\tau = \frac{\sum_1^K 1 - R_k/\max(R_k)}{K-1}, \tag{12}$$

in which $\max(R_k)$ is the gene's maximum expression level across all tissues [37]. We obtained protein-protein interaction data from HuRI [39] and computed each gene's PPI degree, *i.e.*, the total number of unique interaction partners. Also, we obtained 2,220 human metabolic genes involved in one or more curated metabolic pathways from MSigDB release 7.1 [52, 53]. Finally, we converted human-based annotations of gene expression level, tissue specificity, PPI degree, and metabolic genes to the panTro4 assembly using dbNSFP version 4.0 [29, 30].

We obtained a fine-scale chimpanzee genetic map from panMap [35] and converted the data from panTro2 to panTro4 using liftOver. Then, we constructed a map of local recombination rates by averaging the recombination rates from the chimpanzee genetic map with a 1Mb non-overlapping sliding window. Also, we utilized interspecies divergence in the chimpanzee lineage to construct a map of local mutation rates in the panTro4 assembly. To do so, we converted putatively neutral regions defined in [56] from hg19 to panTro4 using liftOver. Then, we computed the density of chimpanzee-specific substitutions in putatively neutral regions using a 100Kb non-overlapping sliding window, which was used as a proxy of local mutation rates.

Estimating the effects of genomic features on the rate of adaptation

We fit the MK regression to one genomic feature at a time, which we named as the simple MK regression. To evaluate if a logarithmic transformation can improve model fitting, we carried out two analyses for each feature. In the first analysis, we standardized the feature by subtracting its mean and dividing by its standard deviation, and then fit the simple MK regression to the standardized feature. In the second analysis, we calculated the logarithm of each feature, standardized the output, and then fit the simple MK regression to the transformed data. Because the logarithm of PPI degree is undefined if the PPI degree is equal to 0, we added a pseudo-count of 1 to the PPI degree before the logarithmic transformation. We computed

the log likelihood of the simple MK regression in each analysis, and used the transformation with a higher log likelihood for each feature throughout this work. We also fit the MK regression to all the features simultaneously, which we named as the multiple MK regression.

Statistical matching analysis

We used statistical matching algorithms to estimate the effects of PPI degree, gene expression level, and tissue specificity after adjusting for potential confounding factors. To estimate the effect of PPI degree, we stratified protein-coding genes into two groups based on PPI degree. The first group consisted of 1,556 genes with 10 or more interaction partners (PPI degree ≥ 10) while the second group consisted of 8,471 genes with 1 or less interaction partners (PPI degree ≤ 1). We calculated ω_α for the two groups of genes separately using equation 10, and calculated the *p*-value of the difference in ω_α between the two gene groups using a two-tailed permutation test with 1,000 resamplings. Also, we used MatchIt to match each gene from the high-PPI group with a gene of similar log expression level and tissue specificity from the low-PPI group [40]. Then, we repeated the calculation of ω_α and *p*-value for the two groups of matched genes.

To estimate the effect of tissue specificity after controlling for gene expression level, we stratified highly expressed genes (mean expression level > 30) into two approximately equal-sized groups based on tissue specificity. The first gene group consisted of 368 highly expressed genes with high tissue specificity while the second group consisted of 369 highly expressed genes with low tissue specificity. Then, we calculated ω_α for the two groups of genes separately using equation 10, and calculated the *p*-value of the difference in ω_α using a two-tailed permutation test with 1,000 resamplings.

Finally, we estimated the effect of gene expression level after controlling for tissue specificity. We stratified tissue-specific genes (tau > 0.85) into two approximately equal-sized groups based on expression level. The first group consisted of 495 tissue-specific genes with high expression level while the second group consisted of 498 tissue-specific genes with low expression level. We calculated ω_α for the two groups of genes separately using equation 10, and calculated the *p*-value of the difference in ω_α using a two-tailed permutation test with 1,000 resamplings.

Tissue and pathway enrichment

We downloaded annotations of tissue-enriched genes from the Human Protein Atlas version 19.3 [36]. To reduce the burden of multiple testing, we focused on analyzing tissues with at least 50 tissue-enriched genes. We then analyzed the enrichment of each set of tissue-enriched genes in the 495 tissue-specific genes with high expression level, in which the 498 tissue-specific genes with low expression level were used as a background gene set. The p-value of each enrichment test was computed using the Fisher's exact test and then adjusted for multiple testing with false-discovery-rate correction. Similarly, we analyzed the enrichment of Reactome pathways in the 495 tissue-specific genes with high expression level using PANTHER [79], in which we used the 498 tissue-specific genes with low expression level as a background gene set.

Acknowledgments

The author acknowledges David McCandlish, Xinru Zhang, Jui-Shan Lin, and Zhihan Liu for useful discussions. This study was supported by start-up funds from Pennsylvania State University to YFH.

Code and Data Availability

The MK regression model and companion training data are available at https://github.com/yifei-lab/MK-regression.

References

[1] Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular Biology and Evolution* 11(5):725–736.

[2] Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* 11(5):715–724.

[3] Yang Z, Nielsen R, Goldman N, Pedersen A (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155(1):431–449.

[4] Hughes AL (2007) Looking for darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* 99(4):364–373.

[5] McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in drosophila. *Nature* 351(6328):652–654.

[6] Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. *Genetics* 158(3):1227–1234.

[7] Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in drosophila. *Nature* 415(6875):1022–1024.

[8] Eyre-Walker A, Keightley PD (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution* 26(9):2097–2108.

[9] Messer PW, Petrov DA (2013) Frequent adaptation and the mcdonald-kreitman test. *Proceedings of the National Academy of Sciences* 110(21):8615–8620.

[10] Galtier N (2016) Adaptive protein evolution in animals and the effective population size hypothesis. *PLOS Genetics* 12(1):1–23.

[11] Haller BC, Messer PW (2017) asymptoticMK: a web-based tool for the asymptotic mcdonald-kreitman test. *G3: Genes, Genomes, Genetics* 7(5):1569–1575.

[12] Uricchio LH, Petrov DA, Enard D (2019) Exploiting selection at linked sites to infer the rate and strength of adaptation. *Nature Ecology & Evolution* 3(6):977–984.

[13] Campos JL, Halligan DL, Haddrill PR, Charlesworth B (2014) The relation between recombination rate and patterns of molecular evolution and variation in drosophila melanogaster. *Molcular Biology and Evolution* 31(4):1010–1028.

[14] Castellano D, Coronado-Zamora M, Campos JL, Barbadilla A, Eyre-Walker A (2016) Adaptive evolution is substantially impeded by Hill-Robertson interference in drosophila. *Molecular Biology and Evolution* 33(2):442–455.

[15] Rousselle M, et al. (2020) Is adaptation limited by mutation? a timescale-dependent effect of genetic diversity on the adaptive substitution rate in animals. *PLOS Genetics* 16(4):1–24.

[16] Marais G, Charlesworth B (2003) Genome evolution: Recombination speeds up adaptive evolution. *Current Biology* 13(2):R68–R70.

[17] Fraïsse C, Puixeu Sala G, Vicoso B (2019) Pleiotropy modulates the efficacy of selection in drosophila melanogaster. *Molecular Biology and Evolution* 36(3):500–515.

[18] Moutinho AF, Trancoso FF, Dutheil JY (2019) The impact of protein architecture on adaptive evolution. *Molecular Biology and Evolution* 36(9):2013–2028.

[19] Afanasyeva A, Bockwoldt M, Cooney CR, I. H, Gossmann TI (2018) Human long intrinsically disordered protein regions are frequent targets of positive selection. *Genome Research* 28(7):975–982.

[20] Enard D, Cai L, Gwennap C, Petrov DA (2016) Viruses are a dominant driver of protein adaptation in mammals. *eLife* 5:e12469.

[21] Luisi P, et al. (2015) Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. *Genome Biology and Evolution* 7(4):1141–1154.

[22] Hvilsom C, et al. (2012) Extensive x-linked adaptive evolution in central chimpanzees. *Proceedings of the National Academy of Sciences* 109(6):2054–2059.

[23] Avila V, Campos JL, Charlesworth B (2015) The effects of sex-biased gene expression and x-linkage on rates of adaptive protein sequence evolution in drosophila. *Biology Letters* 11(4):20150117.

[24] Campos JL, Johnston KJA, Charlesworth B (2018) The effects of sex-biased gene expression and X-linkage on rates of sequence evolution in drosophila. *Molecular Biolology and Evolution* 35(3):655–665.

[25] Pröschel M, Zhang Z, Parsch J (2006) Widespread adaptive evolution of drosophila genes with sex-biased expression. *Genetics* 174(2):893.

[26] Castellano D, Uricchio LH, Munch K, Enard D (2019) Viruses rule over adaptation in conserved human proteins. *bioRxiv.* doi: https://doi.org/10.1101/555060.

[27] Booker TR, Jackson BC, Keightley PD (2017) Detecting positive selection in the genome. *BMC Biology* 15(1):98.

[28] Moutinho AF, Bataillon T, Dutheil JY (2019) Variation of the adaptive substitution rate between species and within genomes. *Evolutionary Ecology* 34:315–338.

[29] Liu X, Jian X, Eric B (2013) dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Human Mutation* 34(9):E2393–E2402.

[30] Liu X, Wu C, Li C, Boerwinkle E (2016) dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site snvs. *Human Mutation* 37(3):235–241.

[31] de Manuel M, et al. (2016) Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* 354(6311):477–481.

[32] Herrero J, et al. (2016) Ensembl comparative genomics resources. *Database (Oxford)* 2016.

[33] Yates AD, et al. (2020) Ensembl 2020. *Nucleic Acids Res* 48(D1):D682–D688.

[34] Tataru P, Mollion M, Glémin S, Bataillon T (2017) Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics* 207(3):1103–1119.

[35] Auton A, et al. (2012) A fine-scale chimpanzee genetic map from population sequencing. *Science* 336(6078):193–198.

[36] Uhlen M, et al. (2015) Tissue-based map of the human proteome. *Science* 347(6220):1260419.

[37] Yanai I, et al. (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21(5):650–659.

[38] Wong WC, et al. (2011) CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* 27(15):2147–2148.

[39] Luck K, et al. (2020) A reference map of the human binary protein interactome. *Nature* 580(7803):402–408.

[40] Ho DE, Imai K, King G, Stuart EA (2011) MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software* 42(8):1–28.

[41] Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* 20(1):110–121.

[42] Hubisz MJ, Pollard KS, Siepel A (2011) PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in Bioinformatics* 12(1):41–51.

[43] Zhang L, Li WH (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular Biology and Evolution* 21(2):236–239.

[44] Zhu J, He F, Hu S, Yu J (2008) On the nature of human housekeeping genes. *Trends in Genetics* 24(10):481–484.

[45] Zhang J, Yang JR (2015) Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics* 16:409–420.

[46] Jassal B, et al. (2020) The reactome pathway knowledgebase. *Nucleic Acids Res* 48(D1):D498–D503.

[47] Schlenke TA, Begun DJ (2003) Natural selection drives drosophila immune system evolution. *Genetics* 164(4):1471–1480.

[48] Nielsen R, et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLOS Biology* 3(6):e170.

[49] Kosiol C, et al. (2008) Patterns of positive selection in six mammalian genomes. *PLoS Genetics* 4(8).

[50] Barreiro LB, Quintana-Murci L (2010) From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews Genetics* 11(1):17–30.

[51] Haerty W, et al. (2007) Evolution in the fast lane: Rapidly evolving sex-related genes in drosophila. *Genetics* 177(3):1321–1335.

[52] Subramanian A, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102(43):15545–15550.

[53] Liberzon A, et al. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27(12):1739–1740.

[54] Huang YF, Gulko B, Siepel A (2017) Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature Genetics* 49(4):618–624.

[55] Huang YF, Siepel A (2019) Estimation of allele-specific fitness effects across human protein-coding sequences and implications for disease. *Genome Research* 29(8):1310–1321.

[56] Huang YF (2020) Unified inference of missense variant effects and gene constraints in the human genome. *PLOS Genetics* 16(7):e1008922.

[57] Pearl J, Glymour M, Jewell N (2016) *Causal Inference in Statistics: A Primer*. (Wiley).

[58] Corbett-Detig RB, Hartl DL, Sackton TB (2015) Natural selection constrains neutral diversity across a wide range of species. *PLOS Biology* 13(4):1–25.

[59] Goldman N, Thorne JL, Jones DT (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149(1):445–458.

[60] Franzosa EA, Xia Y (2009) Structural determinants of protein evolution are context-sensitive at the residue level. *Molecular Biology and Evolution* 26(10):2387–2395.

[61] Liberles DA, et al. (2012) The interface of protein structure, protein biophysics, and molecular evolution. *Protein Science* 21(6):769–785.

[62] Echave J, Spielman SJ, Wilke CO (2016) Causes of evolutionary rate variation among protein sites. *Nature Review Genetics* 17(2):109–121.

[63] Bloom JD, et al. (2005) Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Sciences* 102(3):606.

[64] Bloom JD, Drummond DA, Arnold FH, Wilke CO (2006) Structural determinants of the rate of protein evolution in yeast. *Molecular Biology and Evolution* 23(9):1751–1761.

[65] Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences* 103(15):5869–5874.

[66] Daub JT, Moretti S, Davydov II, Excoffier L, Robinson-Rechavi M (2017) Detection of pathways affected by positive selection in primate lineages ancestral to humans. *Molecular Biology and Evolution* 34(6):1391–1402.

[67] Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA (2007) Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nature Genetics* 39:1140.

[68] Blekhman R, Oshlack A, Chabot AE, Smyth GK, Gilad Y (2008) Gene regulation in primates evolves under tissue-specific selection pressures. *PLOS Genetics* 4(11):1–13.

[69] Blekhman R, et al. (2014) Comparative metabolomics in primates reveals the effects of diet and gene regulatory variation on metabolic divergence. *Scientific Reports* 4(1):5809.

[70] Csilléry K, Rodríguez-Verdugo A, Rellstab C, Guillaume F (2018) Detecting the genomic signal of polygenic adaptation and the role of epistasis in evolution. *Molecular Ecology* 27(3):606–612.

[71] Barghi N, Hermisson J, Schlötterer C (2020) Polygenic adaptation: a unifying framework to understand positive selection. *Nature Reviews Genetics* 21(12):769–781.

[72] Hastie T (1990) *Generalized Additive Models*. (Routledge).

[73] Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61(4):893–903.

[74] Jukes TH, Cantor CR (1969) Evolution of protein molecules in *Mammalian Protein Metabolism*. (Academic Press), pp. 21–132.

[75] Bierne N, Eyre-Walker A (2004) The genomic rate of adaptive amino acid substitution in drosophila. *Molecular Biology and Evolution* 21(7):1350–1360.

[76] Gossmann TI, et al. (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution* 27(8):1822–1832.

[77] Charlesworth B (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research* 63(3):213–227.

[78] Haeussler M, et al. (2019) The UCSC Genome Browser database: 2019 update. *Nucleic Acids Research* 47(D1):D853–D858.

[79] Mi H, et al. (2017) PANTHER version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research* 45(D1):D183–D189.

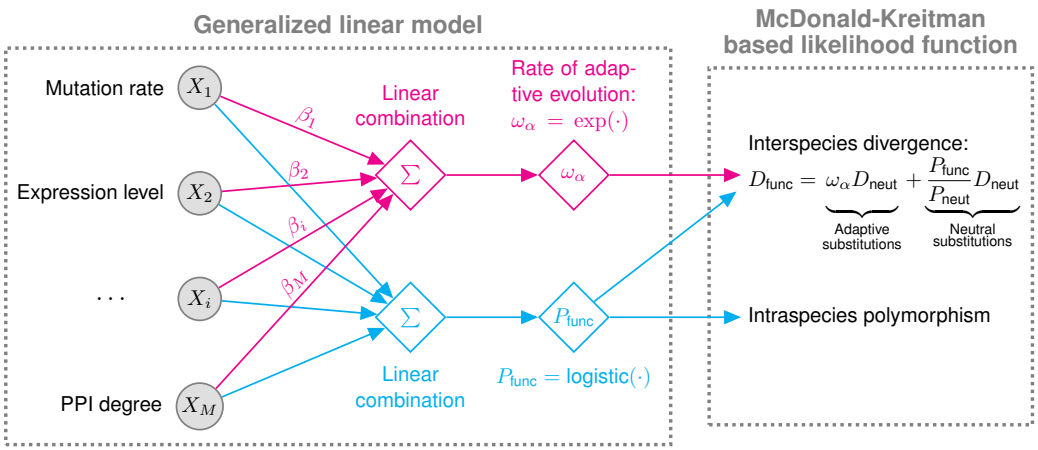

Figure 1: Schematic of the MK regression. The MK regression consists of two components: a generalized linear model and a McDonald-Kreitman based likelihood function. First, we assume that, in a site-wise manner, the rate of adaptive evolution (ω_α) at a functional site is a linear combination of local genomic features followed by an exponential transformation, in which regression coefficient β_i indicates the effect of the i-th feature on adaptive evolution. Similarly, we assume that the probability of observing a SNP (P_{func}) at the same functional site is another linear combination of the same set of genomic features, followed by a logistic transformation. Second, in the McDonald-Kreitman based likelihood function, we combine ω_α and P_{func} at every functional site with two neutral parameters, D_{neut} and P_{neut}, to calculate the probability of observed divergence and polymorphism data given model parameters. D_{neut} and P_{neut} denote the expected number of substitutions and the probability of observing a SNP at a neutral site, respectively. D_{func} denotes the expected number of substitutions at a functional site.

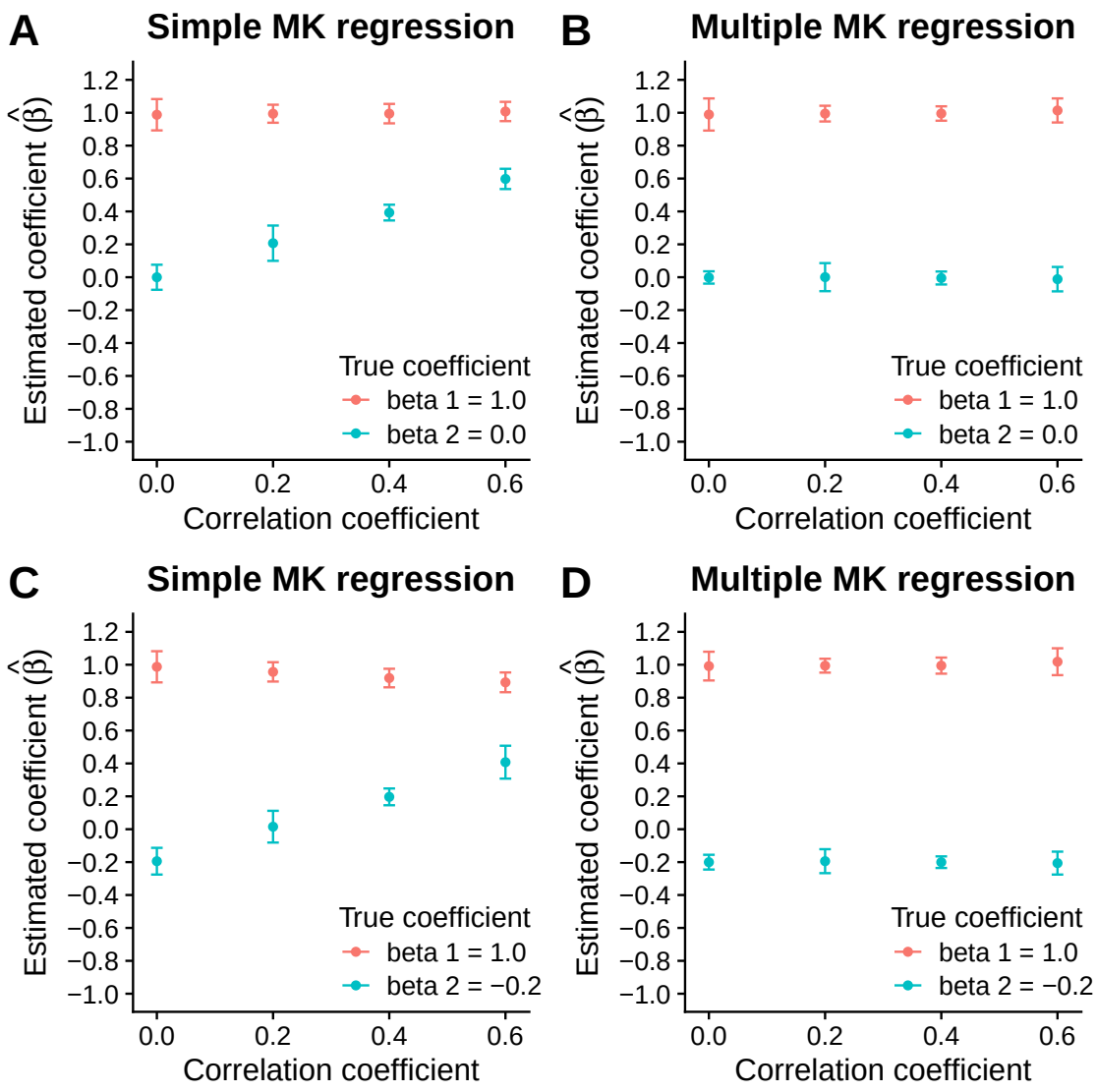

Figure 2: Simulation results. (A) Estimates of regression coefficients in the simple MK regression. The true coefficients are $\beta_1 = 1$ and $\beta_2 = 0$. (B) Estimates of regression coefficients in the multiple MK regression. The true coefficients are $\beta_1 = 1$ and $\beta_2 = 0$. (C) Estimates of regression coefficients in the simple MK regression. The true coefficients are $\beta_1 = 1$ and $\beta_2 = -0.2$. (D) Estimates of regression coefficients in the multiple MK regression. The true coefficients are $\beta_1 = 1$ and $\beta_2 = -0.2$. In each plot, dots and error bars indicate the means and the two-fold standard deviations of estimated coefficients across 10 independent replicates, respectively.

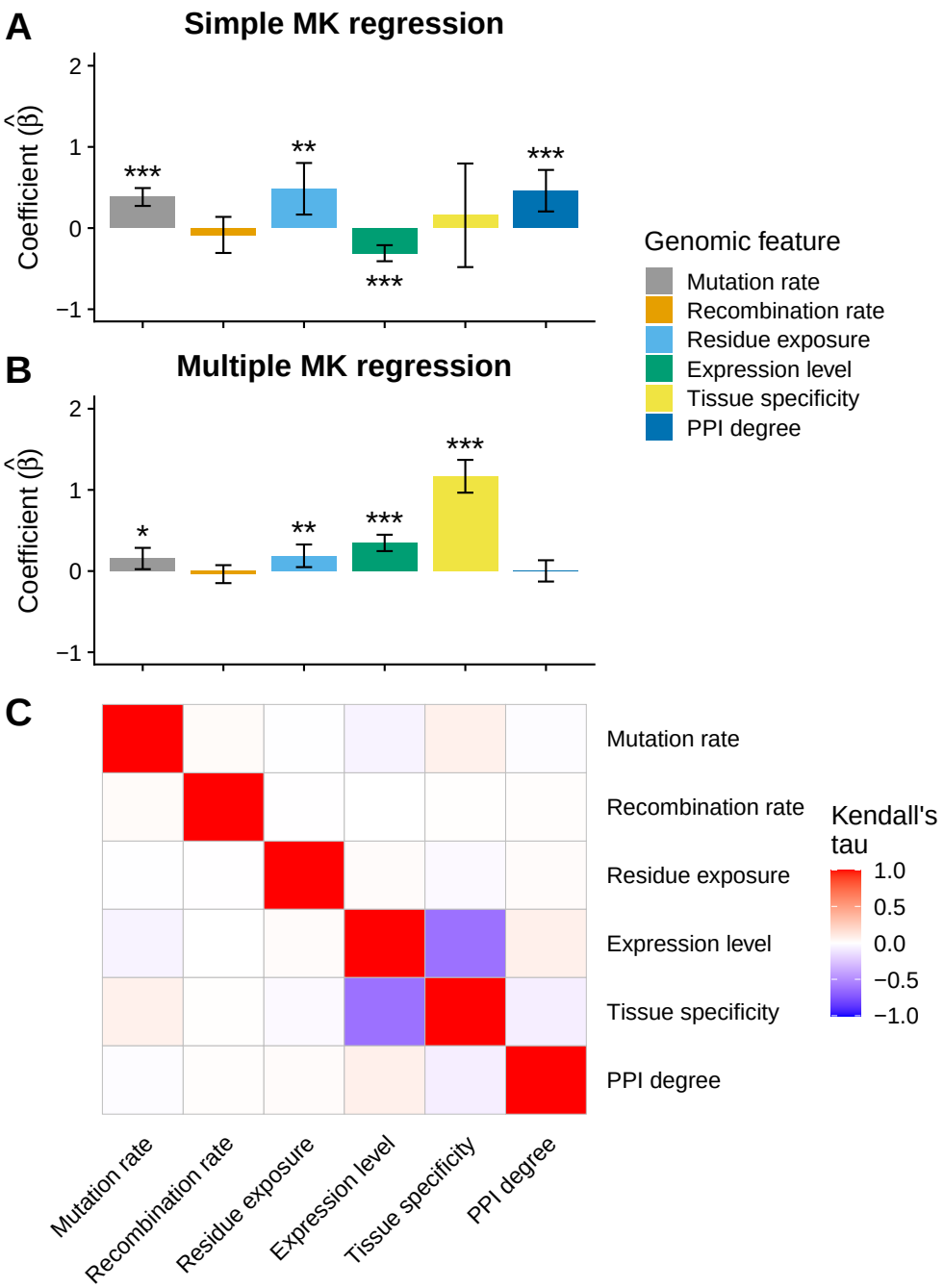

Figure 3: Effects of genomic features on the rate of adaptive evolution. (A) Estimated coefficients of genomic features in the simple MK regression. (B) Estimated coefficients of genomic features in the multiple MK regression. In each bar plot, error bars indicate 95% confidence intervals while one, two, and three asterisks indicate $0.01 \leq p$-value < 0.05, $0.001 \leq p$-value < 0.01, and p-value < 0.001, respectively. (C) Correlations between genomic features.

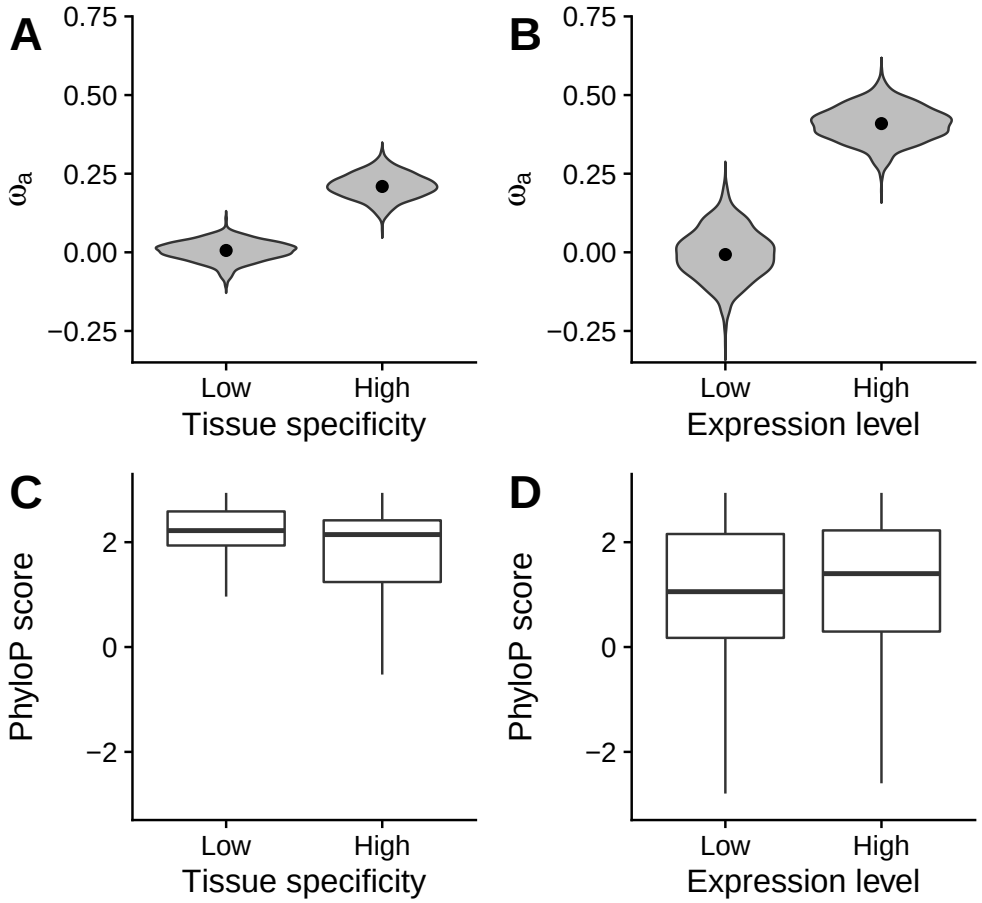

Figure 4: Statistical matching analysis. (A) Estimates of ω_α in 369 highly expressed genes with low tissue specificity and 368 highly expressed genes with high tissue specificity. (B) Estimates of ω_α in 498 tissue-specific genes with low expression level and 495 tissue-specific genes with high expression level. In each violin plot, dots indicate point estimates of ω_α while violins depict the distributions of ω_α from a gene-based bootstrapping analysis with 1,000 resamplings. (C) Distributions of phyloP scores in 369 highly expressed genes with low tissue specificity and 368 highly expressed genes with high tissue specificity. (D) Distributions of phyloP scores in 498 tissue-specific genes with low expression level and 495 tissue-specific genes with high expression level. In each box plot, the bottom, the top, and the middle horizontal bar of the box indicate the first quartile, the third quartile, and the median of phyloP scores, respectively. The whiskers indicate 1.5-fold interquartile ranges.

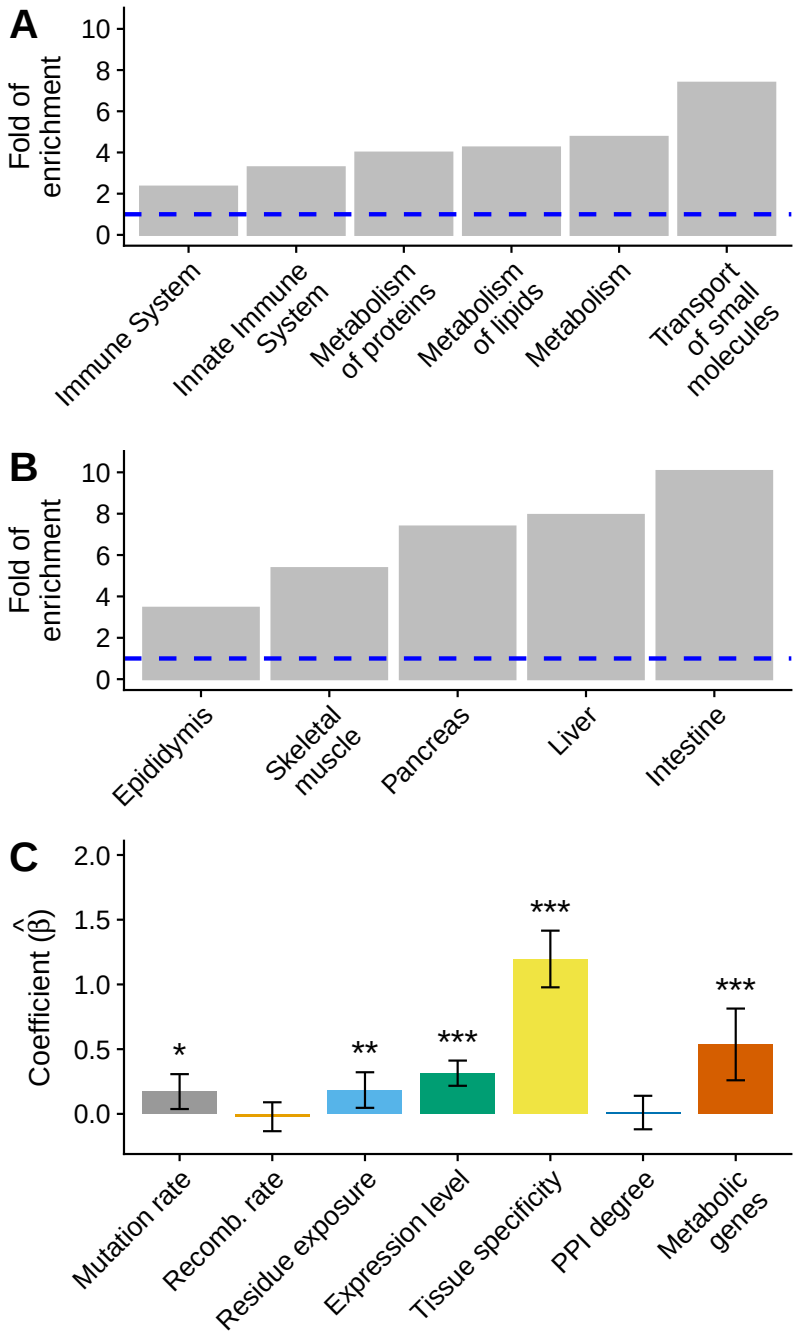

Figure 5: Positive selection in metabolic genes. (A) Enrichment of Reactome pathways in 495 tissue-specific genes with high expression level. (B) Enrichment of tissue types in 495 tissue-specific genes with high expression level. In each enrichment test, 498 tissue-specific genes with low expression level are used as a background gene set. (C) Estimates of regression coefficients in the multiple MK regression. This analysis includes a new binary feature indicating whether each 0D site is located in a metabolic gene. Error bars indicate 95% confidence intervals while one, two, and three asterisks indicate $0.01 \leq p$-value < 0.05, $0.001 \leq p$-value < 0.01, and p-value < 0.001, respectively.