



Published in final edited form as:

Nat Ecol Evol. 2019 June ; 3(6): 977–984. doi:10.1038/s41559-019-0890-6.

Exploiting selection at linked sites to infer the rate and strength of adaptation

Lawrence H. Uricchio^{1,†,*}, Dmitri A. Petrov¹, David Enard^{2,†}

¹Department of Biology, Stanford University, Stanford, CA 94305

²Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721

Abstract

Genomic data encodes past evolutionary events and has the potential to reveal the strength, rate, and biological drivers of adaptation. However, jointly estimating adaptation rate (α) and adaptation strength remains challenging because evolutionary processes such as demography, linkage, and non-neutral polymorphism can confound inference. Here, we exploit the influence of background selection to reduce the fixation rate of weakly-beneficial alleles to jointly infer the strength and rate of adaptation. We develop an MK-based method (ABC-MK) to infer adaptation rate and strength, and estimate $\alpha = 0.135$ in human protein-coding sequences, 72% of which is contributed by weakly-adaptive variants. We show that in this adaptation regime α is reduced $\approx 25\%$ by linkage genome-wide. Moreover, we show that virus-interacting proteins (VIPs) undergo adaptation that is both stronger and nearly twice as frequent as the genome average ($\alpha = 0.224$, 56% due to strongly-beneficial alleles). Our results suggest that while most adaptation in human proteins is weakly-beneficial, adaptation to viruses is often strongly-beneficial. Our method provides a robust framework for estimating adaptation rate and strength across species.

Introduction

The relative importance of selection and drift in driving species' diversification has been a matter of debate since the origins of evolutionary biology. In the earliest formulations of evolutionary theory, natural selection was proposed to be the predominant driver of differences between species [1, 2]. Subsequent theorists argued that random genetic drift could be a more important contributor to species differences [3, 4, 5, 6], with random changes accumulating over evolutionary time due to reproductive isolation between populations. Although it is now clear that natural selection plays a substantial role in both diversification and constraint in many species [7, 8, 9, 10], considerable uncertainty remains about the relative importance of stochastic drift, mutation, selection, and linkage, with no clear consensus among evolutionary geneticists [11, 12, 13, 14]. A better mechanistic

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

[†]To whom correspondence should be addressed: uricchio@berkeley.edu, denard@email.arizona.edu.

*Current address: Department of Integrative Biology, University of California, Berkeley

Competing Interests

The authors declare no competing interests.

understanding of these processes and how they jointly shape genetic diversity could help to resolve old evolutionary puzzles, such as the narrow range of observed genetic diversity across species [15] and the apparently low rate of adaptation in primates [16].

With the exception of rapidly evolving microbial species, most adaptation events occur too slowly to be directly observed over the timescale of a scientific study. Therefore, detailed study of the molecular basis of adaptation has required the development of computational methods to infer adaptation rates (denoted α , defined as the proportion of fixed differences between species that confer fitness benefits) directly from genetic sequence data. Most existing approaches derive from the McDonald-Kreitman (MK) test [7, 17] and related Poisson random field framework [18], both of which use divergence and polymorphism data to infer adaptation rates. Note that a recent approach uses polymorphism data alone to infer the distribution of fitness effects of fixing mutations [19]. The critical idea behind each of these methods is to compare evidence for differentiation at alleles that are likely to have fitness effects (*e.g.*, nonsynonymous alleles that change protein function by altering the amino acid sequence) to alleles that are less likely to have fitness effects (*e.g.*, synonymous alleles that do not change amino acid sequences).

In the classic MK framework, the rate of divergence at putatively functional sites (D_N , often defined as non-synonymous differences within proteins) is compared to putatively neutral diverged sites (D_S , often defined as synonymous differences). Polymorphic sites within both the functional and non-functional class (P_N and P_S , respectively) are used as a background to calibrate the expected rate of divergence under a neutral model. If mutations at functional sites are assumed to be either virtually lethal or neutral, then the ratio $\frac{D_N}{D_S}$ has the same

expected value as $\frac{P_N}{P_S}$ given that virtually lethal mutations contribute to neither P_N nor D_N .

When $\frac{D_N}{D_S}$ exceeds $\frac{P_N}{P_S}$, this is interpreted as evidence of adaptation because sites with functional effects on proteins are over-represented among the fixed differences relative to the neutral expectation. Smith and Eyre-Walker developed a simple equation that uses the same logic as the MK test to estimate adaptation rate α ,

$$\alpha \approx 1 - \frac{D_S P_N}{D_N P_S}, \quad (1)$$

and used this approach to provide evidence for a high rate of adaptation in *Drosophila* [17]. In principle, non-adaptive processes (*i.e.*, processes that do not increase fitness) such as GC-biased gene conversion [20] could also lead to an excess of nonsynonymous fixed differences between species, but only if these processes differentially affect synonymous and nonsynonymous mutations.

Unfortunately, this elegant framework is susceptible to many biases, most notably driven by the presence of weakly deleterious polymorphism in the class P_N . Deleterious polymorphism effectively makes the test overly conservative, because deleterious alleles are unlikely to ever

reach fixation and therefore lead to the overestimation of the expected background rate of substitutions in the functional class. Fay *et al* introduced the idea of including only common polymorphic alleles (*e.g.*, alleles at frequency 15% or greater), which should remove many deleterious alleles [21] – however, this approach has been shown to provide conservative adaptation rate estimates in many contexts [22]. More recently, Messer & Petrov showed that even removing all polymorphism below 50% is insufficient to correct this bias, especially when slightly deleterious mutations are common and the rate of adaptive evolution is high [23]. In order to mitigate this effect, Messer & Petrov introduced the idea of the asymptotic MK test (aMK). In this implementation, $\frac{P_N}{P_S}$ in eqn. 1 is replaced by $\frac{P_N(x)}{P_S(x)}$,

where $P_N(x)$ and $P_S(x)$ are the number of segregating nonsynonymous and synonymous alleles at frequency x , respectively [23]. An exponential curve is fit to the resulting $\alpha(x)$ function, which can be calculated for all values of x in the interval (0,1) for a sample of sequenced chromosomes. The intercept of the best-fit exponential curve at $x = 1$ is a good approximation for α , as it effectively removes all slightly deleterious polymorphism at all frequencies. This approach was shown to be robust to both the underlying distribution of deleterious effects and recent demographic events [23]. aMK has inspired new approaches to inferring adaptation in mitochondrial genes [24] and revealed a high rate of adaptation in proteins interacting with pathogens [25].

While aMK extends the elegant MK framework for estimating adaptation rate, it does not explicitly account for the possibility that beneficial alleles contribute to segregating polymorphism. It is unknown whether aMK is robust to the presence of weakly beneficial alleles, but there is reason to believe that beneficial alleles would be problematic because they are preferentially found at very high frequencies [19], and thus their effect would not be eliminated by the asymptotic procedure. The recent emphasis on adaptation from standing variation [26, 27, 28, 29, 30] and the reported evidence for weakly-beneficial polymorphism in *Drosophila* [31] suggest that robust methods for inferring adaptation strength over longer evolutionary time-scales are needed.

A key limitation of existing MK-based approaches is that they provide estimates of adaptation rate but not adaptation strength, and therefore it is not clear whether weakly beneficial mutations contribute substantially to the fixation process. The underlying processes driving weak and strong adaptation might differ, and the ability to separately estimate rates of weak and strong adaptation could provide insight into the biological drivers of adaptation. We hypothesized that such a method could be developed by exploiting the impact of background selection (BGS) on the fixation rate of weakly-beneficial alleles. BGS removes neutral and weakly-beneficial variation via linkage to deleterious loci [32], while the fixation rate of strongly-adaptive alleles is not substantially affected [33]. Given that the strength of BGS varies widely and predictably across the human genome [34], a method that interrogates the rate of adaptation as a function of BGS might be able to jointly infer the rate and strength of adaptation.

Here, we probe the performance of aMK when weakly-beneficial alleles substantially contribute to segregating polymorphism, and we show that aMK underestimates α in this adaptation regime. We additionally show that when adaptation is weak, true α is predicted to

vary substantially across the genome as a function of the strength of BGS. We exploit this signal of covariation between α and BGS in the weak-adaptation regime to develop an approximate Bayesian computation method, which we call ABC-MK, that separately infers the rate of adaptation for weakly-beneficial and strongly-beneficial alleles. Our approach and aMK rely on similar input data, but we use a model-based fitting procedure that directly accounts for BGS and weakly-beneficial alleles. We apply our method to human genetic data to provide evidence that adaptation in humans is primarily weakly-beneficial and varies as a function of BGS strength. Interestingly, adaptation rate estimates on virus-interacting proteins support a much higher rate of strong adaptation, suggesting that adaptation to viruses is both frequent and strongly fitness-increasing. We address seven potential sources of confounding, and discuss our results in light of recent research on adaptation in humans.

Results

α estimates are conservative for weakly-beneficial selection

The aMK approach is known to converge to the true α at high frequency under the assumption that positively-selected mutations make negligible contributions to the frequency spectrum [23]. This assumption is likely to be met when beneficial alleles confer large fitness benefits, because selective sweeps occur rapidly and beneficial alleles are rarely observed as polymorphic. However, when selection is predominantly weak, attaining a substantial α requires much larger mutation rates for beneficial alleles and longer average transit time to fixation, introducing the possibility that weakly-beneficial alleles will contribute non-negligibly to the frequency spectrum.

To test whether aMK is sensitive to polymorphic weakly-adaptive alleles, we used simulated polymorphism and divergence data to estimate the rate of adaptation using published aMK software [35]. In our simulations, we set the true value of α to 0.2 and vary the contribution of weakly-beneficial alleles and strongly-beneficial alleles to the adaptation process (see Methods & Supplementary Information). When adaptation is due entirely to strongly-adaptive alleles, the estimated value of $\alpha(\hat{\alpha})$ was close to the true value but slightly conservative ($\hat{\alpha} = 0.181 \pm 0.01$; Fig. 1A). As we increased the contribution of weakly-beneficial alleles to α , estimates of α became increasingly conservative ($\hat{\alpha} = 0.144 \pm 0.01$ when $\alpha_W = 0.1$, and $\hat{\alpha} = 0.122 \pm 0.015$ when $\alpha_W = 0.2$; Fig. 1B–C). Removing polymorphism above frequency 0.5 has been suggested as approach to account for potential biases induced by high-frequency derived alleles, which could be mispolarized in real datasets [25]. Restricting to alleles below frequency 0.5 produced similar (but conservative) estimates for all three models ($\hat{\alpha} = 0.14271, 0.14529$, and 0.14264 for $\alpha_W = 0.0, 0.1$ and 0.2 , respectively), likely because the frequency spectrum is not strongly dependent on the rate of weakly-beneficial mutation for low-frequency alleles. Lastly, we performed a much larger parameter sweep across α values and selection coefficients. We find that α estimates become increasingly conservative as the proportion of weakly deleterious alleles increases, and as the strength of selection at beneficial alleles decreases (Fig. S12A & Supplementary Information). Asymptotic-MK estimates of α are only weakly dependent on the distribution of deleterious selection coefficients (Fig. S12).

To better understand why parameter estimates decreased as the proportion of weakly-adaptive alleles increased, we performed analytical calculations of $\alpha(x)$ using diffusion theory [36, 37]. Since we use large sample sizes in our analysis herein, we replace the terms $p_M(x)$ and $p_S(x)$ in $\alpha(x)$ with ${}_x p_M(x)$ and ${}_x p_S(x)$ in our calculations, which trivially asymptote to the same value as the original formulation but are not strongly affected by sample size (see Supplementary Information). We find that the downward bias in estimates of α is due to segregating weakly-adaptive alleles, and removing these alleles from the simulated and calculated $\alpha(x)$ curves would restore the convergence of $\alpha(x)$ to the true α at high frequency (Fig. 1A–C, red curves). In real data, it is not possible to perfectly partition positively-selected and deleterious polymorphic alleles. Hence, in later sections we focus on using the shape of the $\alpha(x)$ curve to infer the strength and rate of adaptation under models that include linkage and complex demography.

Background selection reduces true α when adaptation is weak

We have shown that weakly-beneficial alleles may impact aMK analyses by contributing to segregating polymorphism. This presents an opportunity to study whether aMK estimates vary as a function of background selection (BGS) strength. BGS, the action of linkage between deleterious alleles and neutral alleles, reduces genetic diversity in the human genome [34] and affects neutral divergence rates [38], and is predicted to decrease the fixation probability of weakly-adaptive alleles [33]. Hence, we hypothesized that if adaptation is partially driven by weakly-beneficial alleles in some species, BGS could play a role in modulating adaptation rate across the genome.

To better understand how BGS might affect aMK inference in the presence of weakly-beneficial alleles, we performed analytical calculations and simulations of $\alpha(x)$ with various levels of BGS. We set $\alpha = 0.2$ in the absence of BGS, and then performed simulations while fixing the rate of adaptive mutations and changing the amount of BGS (ranging from $\frac{\pi}{\pi_0} = 0.4$ to 1.0, where π is neutral nucleotide diversity as compared to the neutral diversity in the absence of linked selection, π_0). We find that when adaptation is strong, BGS has a modest effect on $\alpha(x)$ and the true value of α (Fig. 2A&C), mostly driven by an increase in the rate of fixation of deleterious alleles (Fig. S2E). When adaptation is weak, BGS removes a substantial portion of weakly-adaptive alleles and precludes them from fixing, resulting in much stronger dependence of $\alpha(x)$ on BGS and a substantial reduction in the true value of α (Fig. 2B&D and Fig. S2C). Similar to the previous section, estimates of α were conservative across all models, but the underestimation was much more pronounced for weak adaptation (Fig. 2C&D).

Human adaptation rate is shaped by linked selection

Our modeling results show that α is likely to be underestimated when weakly-beneficial alleles contribute substantially to the frequency spectrum, and that background selection may reduce adaptation rate when fitness benefits of adaptive alleles are small. Since BGS is thought to drive broad-scale patterns of diversity across the human genome [34], we hypothesized that directly accounting for the action of BGS on adaptation rate could provide new insights into the evolutionary mechanisms driving adaptation. Moreover, the fact that

weak adaptation is strongly affected by BGS while strong adaptation is not suggests that strong and weak adaptation could be differentiated in genomic data by comparing regions of differing BGS strengths (from $\frac{\pi}{\pi_0} = 0.2$ to $\frac{\pi}{\pi_0} = 1$). We therefore designed an ABC-based method to infer α while accounting for both BGS and weakly-beneficial alleles.

We applied our inference procedure (ABC-MK) to empirical $\alpha(x)$ data computed from human genomes obtained from the Thousand Genomes Project (TGP) for all 661 samples with African ancestry [39]. We find strong posterior support for a substantial component of α driven by weakly-beneficial alleles ($\hat{\alpha}_W = 0.097$; Fig. 3A & see Tab. 1 for area of 95% HPD), as well as posterior support for a smaller component of α from strongly-beneficial alleles $\hat{\alpha}_S = 0.041$). We estimate that the total $\hat{\alpha} = 0.135$, nearly twice the estimate obtained with the same dataset using the original aMK approach ($\hat{\alpha} = 0.076$, see Supplementary Information; we note that while our estimate is similar to previous estimates [40, 23], we use a much larger set of genes in our inference and hence the estimates are not directly comparable). In addition to rates of positive selection, our approach provides estimates of negative selection strength. We find support for mean strength of negative selection of $2Ns \approx -220$ (Fig. S9C), which is consistent with recent studies using large sample sizes [41] and weaker than earlier estimates using small samples [42, 40].

In addition to estimating evolutionary parameters, we sought to better understand how BGS may impact adaptation rate across the genome. We resampled parameter values from our posterior estimates of each parameter, and ran a new set of forward simulations using these parameter values. We then calculated α as a function of BGS in our simulations. We find that α co-varies strongly with BGS, with α in the lowest BGS bins being 33% of α in the highest bins (Fig. 3C). Integrating across the whole genome, our results suggest that human adaptation rate in coding regions is reduced by approximately 25% by BGS (Fig. S9D). To confirm that these model projections are supported by the underlying data, we split the genome into BGS bins and separately estimated adaptation rate in each bin. Although these estimates are substantially noisier than our inference on the full dataset, we find that the rate of adaptation due to weakly-beneficial alleles decreases as a function of BGS strength in accordance with the model predictions (Fig. 3D). In contrast, estimates of the mean strength of negative selection against nonsynonymous mutations did not covary with BGS strength (Fig. S20). Lastly, to validate that our model recapitulates $\alpha(x)$ values that we observe in real data, we also used our independent forward simulations to recompute $\alpha(x)$. We find that our model is in tight agreement with the observed data across the majority of the frequency spectrum. The model and data deviate at high frequency, but both are within the sampling uncertainty (Fig. 3B, gray envelope).

Previous research has shown that virus-interacting proteins (VIPs) have undergone faster rates of adaptation than the genome background [25]. However, the strength of selection acting on these genes is unknown, and given our BGS results it is plausible that the higher rate of adaptation in VIPs is driven by lower overall background selection at VIPs rather than increased selection pressure for adaptation. In contrast, if pathogens have imposed large fitness costs on humans it is possible that VIPs would support both higher adaptation rates

and greater adaptation strength. We ran our method while restricting to an expanded set of 4,066 VIPs for which we had divergence and polymorphism data available. We found evidence for strikingly higher adaptation rates in VIPs than the genome background ($\alpha = 0.224$) and a much larger contribution from strongly-adaptive alleles ($\alpha_S = 0.126$; Fig. 4). The higher α for VIPs cannot be explained by BGS, because VIPs undergo slightly stronger BGS than average genes; the mean BGS strength at VIPs is 0.574, as compared to 0.629 for all genes (in units of π/π_0). Taking α_S as a point estimate for the rate of strongly-beneficial substitutions in VIPs and α_S genome-wide, we estimate that 61% of all strongly-beneficial substitutions occurred in VIPs (Tab. 1). Moreover, we estimate that the posterior probability that α is greater in VIPs than non-VIPs is 99.97%, while the posterior probability that α_S is greater in VIPs is 88.9% (Fig. 4C). Bootstrap samples of non-VIPs (1,000 replicates) never resulted in α_S estimates as high as those obtained from VIPs (Fig. S19). These results are concordant with the $\alpha(x)$ summary statistics for VIPs, which had larger values at high frequency alleles than non-VIPs (Fig. 4D). Interestingly, $\alpha(x)$ is lower for VIPs than the non-VIPs at low frequency, suggesting increased overall levels of conservation among VIPs (see also Fig. S9, where we find support for stronger negative selection against nonsynonymous mutations in VIPs).

Discussion

A long-running debate in evolutionary biology has concerned the relative importance of drift and selection in determining the rate of diversification between species [3, 4, 6, 13]. While previous studies have shown that there is a substantial signal of adaptation in *Drosophila* [17], estimates of adaptation rate in humans are much lower [13]. Here, we extended the classic MK framework to account for weakly-beneficial alleles, and we provided evidence for a large rate of weakly-adaptive mutation in humans. We showed that a state-of-the-art approach to adaptation rate estimation that does not account for beneficial polymorphism provides conservative estimates of α ($\hat{\alpha} = 0.076$ for this data) [23], while our method nearly doubles the estimated human adaptation rate (to $\hat{\alpha} = 0.135$). Most of the adaptation signal that we detect is due to weakly-beneficial alleles. Interestingly, virus-interacting proteins supported a much higher rate of adaptation than the genome background ($\hat{\alpha} = 0.226$), especially for strongly-beneficial substitutions $\hat{\alpha}_S = 0.126$ as compared to $\hat{\alpha}_S = 0.041$ genome-wide). Our results provide an evolutionary mechanism that partially explains the apparently low observed rate of human adaptation in previous studies, and extends the support for viruses as a major driver of adaptation in humans [25].

It has long been known that recombination could in principle affect the evolutionary trajectories of both beneficial and deleterious alleles [43, 44, 33], and studies in *Drosophila* [45, 46] and dogs [47] have provided evidence for the effect of recombination on divergence and load. Despite the expectation that recombination could have a strong effect on adaptation in humans, studies have differed on how recombination affects human divergence and polymorphism. One human genomic study explored the ratio $\frac{D_N}{D_S}$ as a function of recombination rate, and found no evidence for an effect of recombination on divergence rate

[48]. Our results may partially explain why $\frac{D_N}{D_S}$ does not fully capture the effect of recombination on divergence in humans. As BGS increases in strength, the rate of accumulation of deleterious alleles increases, while the rate of fixation of weakly-adaptive alleles decreases. The two effects partially offset each other, which should reduce the sensitivity of $\frac{D_N}{D_S}$ as a tool to detect the effect of recombination on divergence. A more recent study provided evidence that recombination affects the accumulation of deleterious polymorphic alleles [49], but did not provide detailed information about the effect of recombination on adaptation. Our results are consistent with the idea that weakly deleterious alleles are predicted to segregate at higher frequencies in regions under strong BGS, and we additionally show that BGS affects the accumulation of weakly-beneficial alleles in humans.

While classic MK approaches estimate only the rate of adaptation, our method extends the MK-framework to provide information about both the rate and strength of selection. Previous approaches to estimating the strength of adaptation have focused on the dip in diversity near sweeping alleles [45, 50, 51, 52, 31] or have directly inferred the DFE from the frequency spectrum [19] – our approach capitalizes on an orthogonal signal of the reduction in fixation rate of weakly-beneficial alleles induced by selection at linked sites. We developed an ABC method to capture this signal, but less computationally intensive methods could also be used - for example, the original aMK approach could be applied in bins of BGS strength. If a substantial proportion of adaptation is due to weakly-beneficial alleles, such an analysis should result in a strong correlation between BGS strength and (potentially conservative) α estimates. However, it should be noted that cryptic covariation between gene functions (such as VIPs) and BGS strength could confound such inferences.

We supposed that the main effects of linked selection in humans were due to background selection, but in principle genetic draft could drive similar patterns. Draft is expected to substantially reduce genetic diversity when sweeps occur frequently, and can impede the fixation of linked beneficial alleles [53, 54]. Previous work has also shown that strong draft can alter the fixation rate and frequency spectra of neutral and deleterious alleles [23]. We performed simulations of strong draft in 1MB flanking sequences surrounding a gene evolving under natural selection and tested the magnitude of the deviation from theoretical predictions under a model of background selection alone. Consistent with previous work, we observe that draft increases the fixation rate of deleterious alleles and thereby decreases α [23]. However, the effect on $\alpha(x)$ is only modest at the frequencies that we use in our inference procedure (*i.e.*, below 75%), even when the strength and rate of positive selection are much larger than we and others have inferred in humans (although there is a modest deviation around 75% frequency, the highest frequency we use in our inference; Fig. S4C&D). This implies that draft due to selected sites outside genes would have to be much stronger than draft due to positive selection inside exons in order to drive the effects that we infer in the human genome. We note that it is likely that in species undergoing both strong, frequent sweeps and BGS (*e.g.*, *Drosophila* – see [31]), draft will contribute to the removal of weakly-beneficial polymorphism.

Selection has left many imprints on the human genome, with studies reporting signatures of selective sweeps [52], soft sweeps [29], background selection [34], negative selection [40, 42], and polygenic adaptation [28]. Still, considerable uncertainty remains about the relative importance of these evolutionary mechanisms, especially as concerns the rate and strength of positive selection. Recent work has suggested that the contrasting adaptation rate estimates of previous studies [51, 52] can be reconciled by arguing that most adaptation signals in humans are consistent with adaptation from standing variation [29]. Our results show that the frequency spectra and patterns of divergence are also consistent with the idea that many adaptive alleles segregate much longer than is expected for a classic sweep, and hence also help to reconcile the results of previous studies.

In addition to determining the rate, strength, and mechanisms of adaptation, there is an ongoing effort to find the biological processes most important for driving adaptation. Previous work has shown that viruses are a critical driver of adaptation in mammals [25], but the strength of the fitness advantages associated with resistance to (or tolerance of) infection remain unclear. Our approach clarifies that strongly-adaptive fixed differences are also approximately three-fold enriched in virus-interacting proteins relative to non-VIPs. In contrast, weak adaptation rate was not substantially different between VIPs and non-VIPs, suggesting that weak adaptation may proceed through mechanisms that are shared across proteins regardless of function (for example, optimization of stability). While we have focused on VIPs here due to the expected fitness burdens associated with infection, in future research our approach could be used to investigate adaptation in any group of genes, or extended to partition genes into strong and weak adaptation classes.

The model that we fit to human data does an excellent job of recapitulating the observed patterns in the Thousand Genomes Project data, but we were concerned that several possible confounding factors could influence our results. We showed that seven confounding factors (ancestral mispolarization [55], demographic model misspecification [56, 57], BGS model misspecification, covariation of BGS and sequence conservation, GC-biased gene conversion [20], selection on synonymous alleles [58], and misspecification of strongly- and/or weakly-beneficial selection coefficients) are unlikely to substantially influence the results (see Supplementary Information), but it should be noted that the adaptive process in our model is exceedingly simple, and it is very likely that the evolutionary processes driving diversification are much more complex. We supposed that adaptation proceeds in two categories, weak and strong selection, each of which is described by a single selection coefficient. In reality, adaptive alleles are likely to have selection coefficients drawn from a broad distribution, and adaptation is likely to proceed by a variety of mechanisms, including sweeps [52], polygenic adaptation [28], and selection from standing variation [29]. While our results show that BGS shapes adaptation rate across the genome, our method does not differentiate among adaptation mechanisms. We expect that future research will further clarify the relative importance of various selection mechanisms to shaping genomic patterns of diversity in the genomes of humans and other organisms [59, 10].

Our method is flexible in that it could be applied to any species for which both divergence/polymorphism data and estimates of background selection strength are available. As with the original aMK approach, we showed that the α estimates we obtained are not highly sensitive

to recent demographic uncertainty. Our approach may therefore be effective for providing more accurate estimates of adaptation rate in non-model species. Despite recent advances, the evolutionary mechanisms that shape genetic diversity across species (which could include linked selection, population size, and/or population demography) remain the subject of debate [15, 11, 12]. Future work using and extending our method, which accounts for the effect of weakly-beneficial alleles on adaptation rate estimates, could help to resolve this open question.

Methods

Divergence and polymorphism data

We retrieved the number of polymorphic sites and their allele frequencies in human coding sequences as well as the number of human-specific fixed substitutions in coding sequences since divergence with chimpanzees. Fixed substitutions were identified by parsimony based on alignments of human (hg19 assembly), chimpanzee (panTro4 assembly) and orangutan (ponAbe2 assembly) coding sequences. Human coding sequences from Ensembl v73 [60] were blatted [61] on the panTro4 and ponAbe2 assemblies and the best corresponding hits were blatted back on the hg19 human assembly to finally identify human-chimp-orangutan best reciprocal orthologous hits. We used the Blatfine option to ensure that even short exons at the edge of coding sequences would be included in the hits. We further used a Blat protein -minIdentity threshold of 60%. The corresponding human, chimp and orangutan coding sequences were then aligned with PRANKs coding sequence evolution model [62] after codons containing undefined positions were removed.

For each human coding gene in Ensembl we considered all possible protein-coding isoforms and aligned separately each isoform between human, chimp and orangutan. The numbers of polymorphic or divergent sites are therefore the numbers over all possible isoforms of a human gene (however, the same polymorphic or divergent site present in multiple isoforms was counted only once). If a polymorphic or divergent site was synonymous in an isoform but non-synonymous in another isoform, we counted it as a single non-synonymous polymorphic or divergent site. Only fixed divergent sites were included, meaning that substitutions still polymorphic in humans were not counted as divergent.

The derived allele frequency of polymorphic sites herein corresponds to the frequency across all African populations from the Thousand Genomes Project phase 3 (TGP), which comprises 661 individuals spread across seven different subpopulations [39]. Allele frequencies were extracted from vcf files provided by the TGP for the phase 3 data. In total, 17,740 human-chimp-orangutan orthologs were included in the analysis. Supplemental Data Table S1 provides the number of synonymous and non-synonymous polymorphic or divergent sites for each of these 17,740 orthologs, as well as the allelic frequencies of the polymorphic sites. Polymorphic sites were counted only if they overlapped those parts of human coding sequences that were aligned with chimp and orangutan coding sequences. The ancestral and derived allele frequencies were based on the ancestral alleles inferred by the TGP phase 3 and available in the previously mentioned vcf files [39].

Model-based simulations and calculations

We tested the robustness of the aMK approach to the presence of weakly-beneficial alleles using simulation and theory. We simulated simultaneous negative and positive selection in coding sequences using model-based forward simulations under a range of scenarios [63, 64]. We supposed that nonsynonymous alleles are under selection, while synonymous alleles are neutral. In each simulation, we set $\alpha = \alpha_W + \alpha_S = 0.2$, where α_W is the component of α due to weakly-beneficial mutations ($2Ns = 10$) and α_S represents strongly-beneficial alleles ($2Ns = 500$). Note that α is not treated as a parameter in the analyses herein; we use analytical theory to calculate the mutation rates for deleterious alleles and advantageous alleles that result in the desired α , meaning that α is a model output and not a model input. We drew deleterious selection coefficients from a Gamma distribution inferred from human sequence data [40], and we varied α_W from 0 to 0.2. We used the simulated allele frequency spectra and fixed differences to calculate the $\alpha(x)$ summary statistics. Results of these simulations are provided in Fig. 1 and additional simulation details are included in the Supplementary Information.

We also performed analytical calculations under the same evolutionary model using results from diffusion theory. These calculations are described in the Supplementary Information (see the sections entitled “Analytical approximation to $\alpha(x)$ ” and “Background selection & adaptive divergence”). Software to perform these calculations is available at <https://github.com/uricchio/mktest>.

Using ABC-MK to infer adaptation rate and strength

We developed an approximate Bayesian computation (ABC) approach for estimating α_W and α_S in the presence of BGS and complex human demography [65]. We sampled parameters from prior distributions corresponding to the shape and scale of deleterious selection coefficients (assumed to be Gamma-distributed) and the rate of mutation of weakly and strongly-beneficial mutations. We performed forward simulations [63, 64] of simultaneous negative and positive selection at a coding locus under a demographic model inferred from NHLBI Exome project African American samples [66] with varying levels of background selection from $\pi/\pi_0 = 0.2$ to $\pi/\pi_0 = 1.0$ and the sampled parameter values. We then calculated $\alpha(x)$ using this simulated data, sampling alleles from the simulations such that the distribution of BGS values in the simulation matches the distribution in the empirical data as calculated by a previous study [34]. We used $\alpha(x)$ values at a subset of frequencies \times as summary statistics in ABC (specifically, at derived allele counts 1, 2, 5, 10, 20, 50, 100, 200, 500, and 1000 in a sample of 1322 chromosomes). To improve efficiency, we employed a resampling-based approach that allows us to query many parameter values using the same set of forward simulations (see Supplementary Information).

We tested our approach by estimating parameter values (population scaled mutation rates θ_S , θ_W , and the parameters of a Gamma distribution controlling negative selection strength) and quantities of interest (α_W , α_S , α) from simulated data. We found that the method produces high-accuracy estimates for most inferred parameters and α values (including α_W , α_S , and total α – Fig. S6). Some parameter values (particularly those corresponding to the distribution of fitness effects (DFE) over deleterious alleles and mutation rates of beneficial

alleles) were somewhat noisily inferred. We found that α estimates were not very sensitive to various types of model misspecification (See Supplementary Information – Robustness analyses), but α_W and α_S are modestly affected by misspecification of the demographic model or the DFE of alleles driving BGS. We call our approach ABC-MK.

Data & code availability

Supplemental Data Table S1 is provided on the publisher's website. The data and code that we used to parameterize our model are freely available online at <https://github.com/uricchio/mktest>. Columns in Supplemental Data Table S1 are as follows: First column – Ensembl coding gene ID. Second column – number of non-synonymous polymorphic sites. Third column – respective derived allele frequencies of these sites separated by commas. Fourth column – number of synonymous polymorphic sites. Fifth column – respective frequencies derived allele frequencies of these sites. Sixth column – number of fixed non-synonymous substitutions on the human branch. Seventh column – number of fixed synonymous substitutions on the human branch.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Philipp Messer, Raul Torres, Ying Zhen, Christian Huber, Kirk Lohmueller, Zachary Szpiech, Adam Siepel, Yifei Huang, and Hussein Hijazi for helpful comments on the manuscript. We thank Alan Aw, Noah Rosenberg, and members of the Rosenberg lab for helpful discussions. LHU was partially supported by and IRACDA fellowship through NIGMS grant K12GM088033, as well as NIH R01 HG005855 and NSF DBI-1458059 (each to Noah Rosenberg). We thank the Stanford/SJSU IRACDA program for support.

References

- [1]. Darwin C On the origin of species. Murray; 1859.
- [2]. Wallace AR. Darwinism: an exposition of the theory of natural selection with some of its applications. MacMillan & Co; 1889.
- [3]. Wright S On the roles of directed and random changes in gene frequency in the genetics of populations. *Evolution*. 1948;2(4):279–294. [PubMed: 18104586]
- [4]. Kimura M, et al. Evolutionary rate at the molecular level. *Nature*. 1968;217(5129):624–626. [PubMed: 5637732]
- [5]. Ohta T Slightly deleterious mutant substitutions in evolution. *Nature*. 1973;246(5428):96. [PubMed: 4585855]
- [6]. Kimura M Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*. 1977;267(5608):275. [PubMed: 865622]
- [7]. McDonald JH, Kreitman M. Adaptive protein evolution at the ADH locus in *Drosophila*. *Nature*. 1991;351(6328):652. [PubMed: 1904993]
- [8]. Fay JC, Wyckoff GJ, Wu CI. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature*. 2002;415(6875):1024. [PubMed: 11875569]
- [9]. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*. 2010;20(1):110–121. [PubMed: 19858363]
- [10]. Charlesworth B, Charlesworth D. Neutral variation in the context of selection. *Molecular Biology and Evolution*. 2018;35(6):1359–1361. [PubMed: 29659981]
- [11]. Corbett-Detig RB, Hartl DL, Sackton TB. Natural selection constrains neutral diversity across a wide range of species. *PLoS Biology*. 2015;13(4):e1002112. [PubMed: 25859758]

- [12]. Coop G Does linked selection explain the narrow range of genetic diversity across species? bioRxiv. 2016;doi:10.1101/042598.
- [13]. Kern AD, Hahn MW. The neutral theory in light of natural selection. *Molecular Biology and Evolution*. 2018;35:1366–1371. [PubMed: 29722831]
- [14]. Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, et al. The importance of the neutral theory in 1968 and 50 years on: a response to Kern and Hahn 2018. *Evolution*. 2018;73:111–114. [PubMed: 30460993]
- [15]. Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, et al. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biology*. 2012;10(9):e1001388. [PubMed: 22984349]
- [16]. Galtier N Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genetics*. 2016;12(1):e1005774. [PubMed: 26752180]
- [17]. Smith NG, Eyre-Walker A. Adaptive protein evolution in *Drosophila*. *Nature*. 2002;415(6875):1022. [PubMed: 11875568]
- [18]. Sawyer SA, Hartl DL. Population genetics of polymorphism and divergence. *Genetics*. 1992;132(4):1161–1176. [PubMed: 1459433]
- [19]. Tataru P, Mollion M, Glémin S, Bataillon T. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics*. 2017;207(3):1103–1119. [PubMed: 28951530]
- [20]. Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, et al. Detecting positive selection within genomes: the problem of biased gene conversion. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. 2010;365(1552):2571–2580. [PubMed: 20643747]
- [21]. Fay JC, Wyckoff GJ, Wu CI. Positive and negative selection on the human genome. *Genetics*. 2001;158(3):1227–1234. [PubMed: 11454770]
- [22]. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution*. 2009;26(9):2097–2108. [PubMed: 19535738]
- [23]. Messer PW, Petrov DA. Frequent adaptation and the McDonald–Kreitman test. *Proceedings of the National Academy of Sciences*. 2013;110(21):8615–8620.
- [24]. James JE, Piganeau G, Eyre-Walker A. The rate of adaptive evolution in animal mitochondria. *Molecular Ecology*. 2016;25(1):67–78. [PubMed: 26578312]
- [25]. Enard D, Cai L, Gwennap C, Petrov DA. Viruses are a dominant driver of protein adaptation in mammals. *eLife*. 2016;5:e12469. [PubMed: 27187613]
- [26]. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*. 2010;20(4):R208–R215. [PubMed: 20178769]
- [27]. Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*. 2013;28(11):659–669. [PubMed: 24075201]
- [28]. Berg JJ, Coop G. A population genetic signal of polygenic adaptation. *PLoS Genetics*. 2014;10(8):e1004412. [PubMed: 25102153]
- [29]. Schrider DR, Kern AD. Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular Biology and Evolution*. 2017;34(8):1863–1877. [PubMed: 28482049]
- [30]. Uricchio LH, Kitano HC, Gusev A, Zaitlen NA. An evolutionary compass for detecting signals of polygenic selection and mutational bias. *Evolution Letters*. 2019;3(1):69–79. [PubMed: 30788143]
- [31]. Elyashiv E, Sattath S, Hu TT, Strutsosky A, McVicker G, Andolfatto P, et al. A genomic map of the effects of linked selection in *Drosophila*. *PLoS Genetics*. 2016;12(8):e1006130. [PubMed: 27536991]
- [32]. Charlesworth B, Morgan M, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. *Genetics*. 1993;134(4):1289–1303. [PubMed: 8375663]
- [33]. Barton NH. Linkage and the limits to natural selection. *Genetics*. 1995;140(2):821–841. [PubMed: 7498757]

- [34]. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genetics*. 2009;5(5):e1000471. [PubMed: 19424416]
- [35]. Haller BC, Messer PW. asymptoticmk: A web-based tool for the asymptotic McDonald–Kreitman test. *G3: Genes, Genomes, Genetics*. 2017;7(5):1569–1575.
- [36]. Evans SN, Shvets Y, Slatkin M. Non-equilibrium theory of the allele frequency spectrum. *Theoretical Population Biology*. 2007;71(1):109–119. [PubMed: 16887160]
- [37]. Kimura M. Diffusion models in population genetics. *Journal of Applied Probability*. 1964;1(2):177–232.
- [38]. Phung TN, Huber CD, Lohmueller KE. Determining the effect of natural selection on linked neutral divergence across species. *PLoS Genetics*. 2016;12(8):e1006199. [PubMed: 27508305]
- [39]. Consortium TGP, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68. [PubMed: 26432245]
- [40]. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* 2008;4(5):e1000083. [PubMed: 18516229]
- [41]. Kim BY, Huber CD, Lohmueller KE. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*. 2017;206(1):345–361. [PubMed: 28249985]
- [42]. Eyre-Walker A, Woolfit M, Phelps T. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*. 2006;173(2):891–900. [PubMed: 16547091]
- [43]. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. *Genetics Research*. 1966;8(3):269–294.
- [44]. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. *Genetics Research*. 1974;23(1):23–35.
- [45]. Macpherson JM, Sella G, Davis JC, Petrov DA. Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. *Genetics*. 2007;177(4):2083–2099. [PubMed: 18073425]
- [46]. Castellano D, Coronado-Zamora M, Campos JL, Barbadilla A, Eyre-Walker A. Adaptive evolution is substantially impeded by Hill–Robertson interference in drosophila. *Molecular Biology and Evolution*. 2015;33(2):442–455. [PubMed: 26494843]
- [47]. Marsden CD, Ortega-Del Vecchyo D, OBrien DP, Taylor JF, Ramirez O, Vilà C, et al. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proceedings of the National Academy of Sciences*. 2016;113(1):152–157.
- [48]. Bullaughey KL, Przeworski M, Coop G. No effect of recombination on the efficacy of natural selection in primates. *Genome Research*. 2008;18:544–554. [PubMed: 18199888]
- [49]. Hussin JG, Hodgkinson A, Idaghdour Y, Grenier JC, Goulet JP, Gbeha E, et al. Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nature Genetics*. 2015;47(4):400. [PubMed: 25685891]
- [50]. Jensen JD, Thornton KR, Andolfatto P. An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. *PLoS Genetics*. 2008;4(9):e1000198. [PubMed: 18802463]
- [51]. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. *Science*. 2011;331(6019):920–924. [PubMed: 21330547]
- [52]. Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. *Genome Research*. 2014;24(6):885–895. [PubMed: 24619126]
- [53]. Comeron JM, Kreitman M. Population, evolutionary and genomic consequences of interference selection. *Genetics*. 2002;161(1):389–410. [PubMed: 12019253]
- [54]. Uricchio LH, Hernandez RD. Robust forward simulations of recurrent hitchhiking. *Genetics*. 2014;197(1):221–236. [PubMed: 24561480]
- [55]. Hernandez RD, Williamson SH, Bustamante CD. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Molecular Biology and Evolution*. 2007;24(8):1792–1800. [PubMed: 17545186]

- [56]. Ewing GB, Jensen JD. The consequences of not accounting for background selection in demographic inference. *Molecular Ecology*. 2016;25(1):135–141. [PubMed: 26394805]
- [57]. Torres R, Szpiech ZA, Hernandez RD. Human demographic history has amplified the effects of background selection across the genome. *PLoS Genetics*. 2018;14(6):e1007387. [PubMed: 29912945]
- [58]. Huang YF, Siepel A. Estimation of allele-specific fitness effects across human protein-coding sequences and implications for disease. *bioRxiv*. 2018;doi:10.1101/441337.
- [59]. Huber CD, Kim BY, Marsden CD, Lohmueller KE. Determining the factors driving selective effects of new nonsynonymous mutations. *Proceedings of the National Academy of Sciences*. 2017;114(17):4465–4470.
- [60]. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. *Nucleic Acids Research*. 2015;44(D1):D710–D716. [PubMed: 26687719]
- [61]. Kent WJ. BLAT – the BLAST-like alignment tool. *Genome Research*. 2002;12(4):656–664. [PubMed: 11932250]
- [62]. Löytynoja A, Goldman N. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics*. 2010;11(1):579. [PubMed: 21110866]
- [63]. Hernandez RD, Uricchio LH. SFS_CODE: more efficient and flexible forward simulations. *bioRxiv*. 2015;doi:10.1101/025064.
- [64]. Uricchio LH, Torres R, Witte JS, Hernandez RD. Population genetic simulations of complex phenotypes with implications for rare variant association tests. *Genetic Epidemiology*. 2015;39(1):35–44. [PubMed: 25417809]
- [65]. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. *Genetics*. 2002;162(4):2025–2035. [PubMed: 12524368]
- [66]. Tennessen JA, Bigham AW, OConnor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012;337(6090):64–69. [PubMed: 22604720]

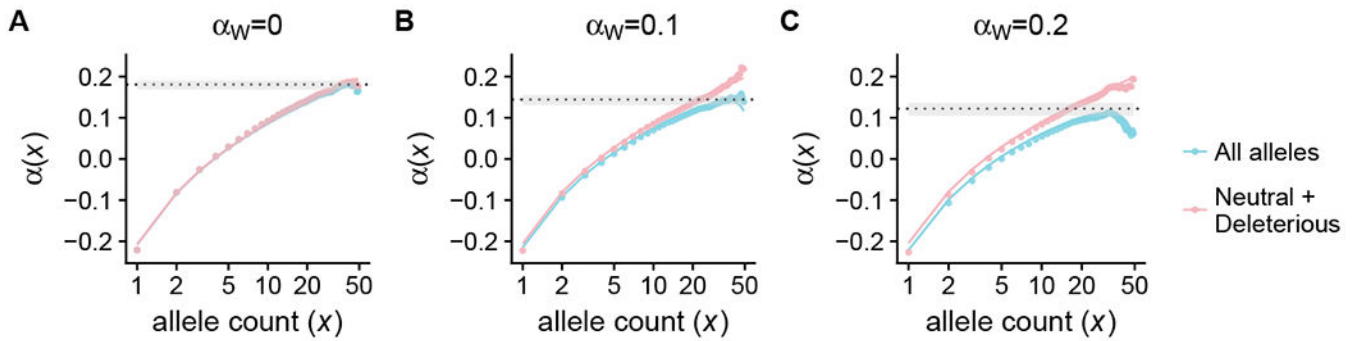


Figure 1: aMK estimates as a function of adaptation strength.

A-C: We plot $\alpha(x)$ as a function of derived allele count x in a sample of 50 chromosomes. The true value of $\alpha = 0.2$ in each panel, with varying contributions from weakly ($2Ns = 10$) and strongly-adaptive alleles ($2Ns = 500$). The solid lines show the results of our analytical approximation (eqn. 11 in the Supplemental Information), while the points show the value of $\alpha(x)$ from forward simulations. The blue points and curves show the calculation as applied to all polymorphic loci, while in the pink points and curves we have removed positively selected alleles from the calculation. The dotted line shows the estimated value of α from the simulated data using existing asymptotic-MK methods [23, 35], while the gray bars show the 95% confidence interval around the estimate.

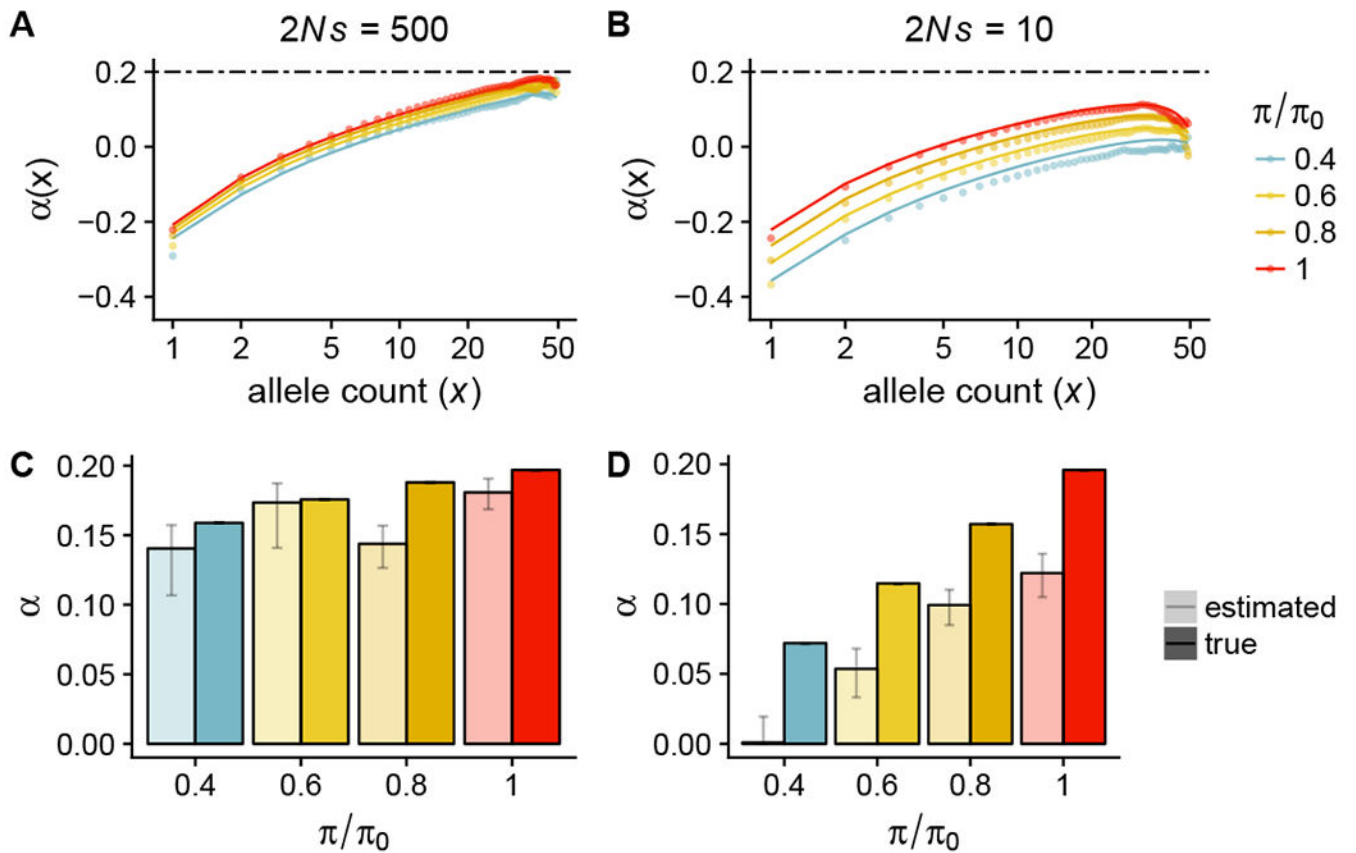


Figure 2: The effect of BGS on α .

A-B: $\alpha(x)$ is plotted as a function of derived allele count x for various background selection (π/π_0) values. In A, adaptive alleles are strongly-beneficial ($2Ns = 500$), while in B they are weakly-beneficial ($2Ns = 10$). The lines represent analytical approximations, while the points represent the results of stochastic simulations. The dashed lines at $\alpha = 0.2$ represent the true rate of adaptation in the absence of BGS. C-D: True (dark colors) and estimated (light colors) α for each of the corresponding models in A-B. Panel C corresponds to strong adaptation ($2Ns = 500$) while D corresponds to weak adaptation ($2Ns = 10$). Estimates of α were made using existing asymptotic-MK software [35], and the error bars correspond to 95% confidence intervals reported by the software. For each parameter combination, we used 2×10^5 independent simulations of 10^3 coding base pairs each.

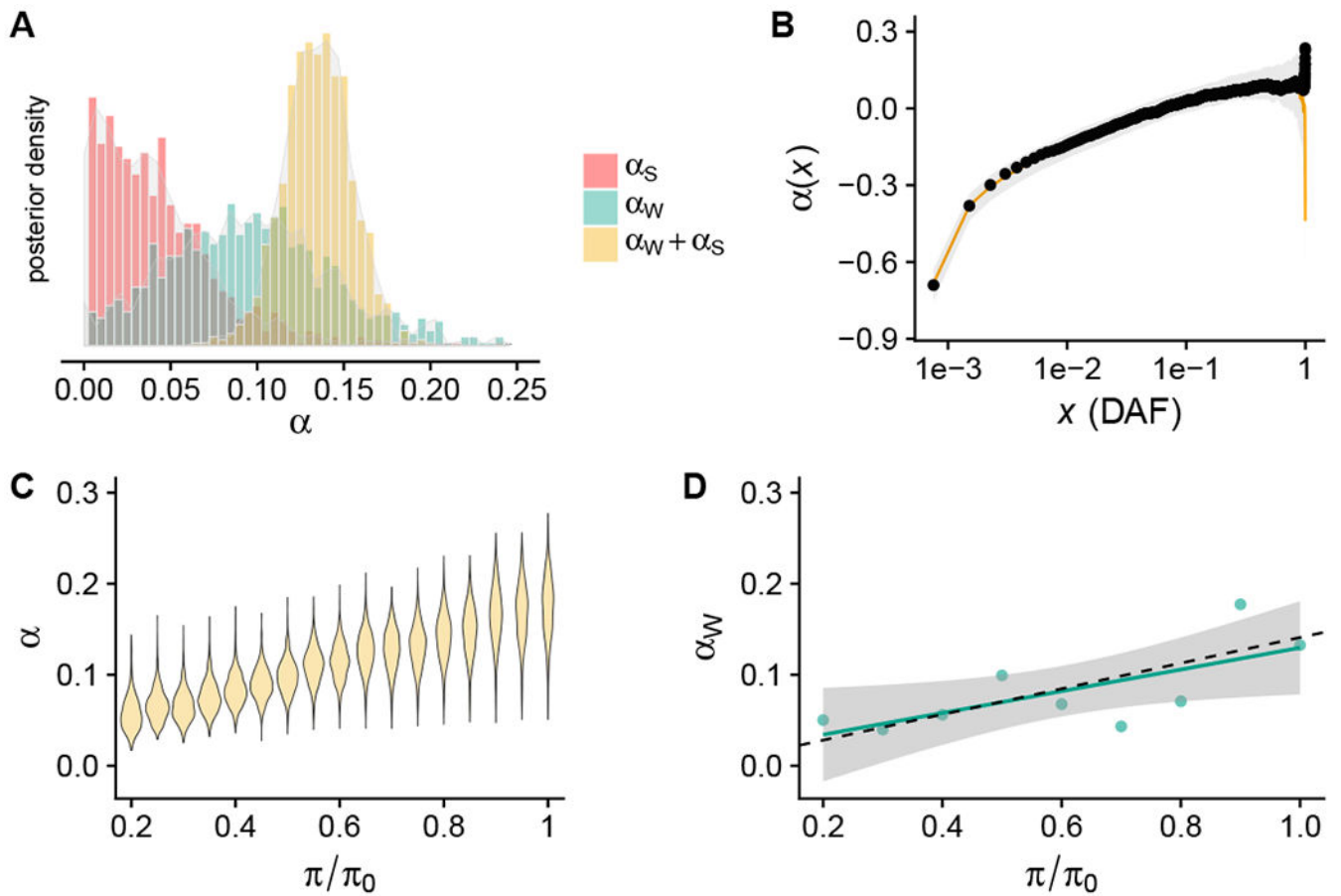


Figure 3: Adaptation rate and strength estimates for human genomic data

A: Posterior distribution of α_W , α_S , and $\alpha = \alpha_S + \alpha_W$ as inferred by applying our ABC approach to 661 samples of African ancestry from the TGP phase 3. B: $\alpha(x)$ as a function of derived allele frequency (DAF) for genomic data (black points) plotted along with the mean posterior estimate from our model (orange line) and 99% confidence interval (gray envelope), as obtained by an independent set of simulations using the posterior parameter estimates. C: Inferred posterior distribution of α as a function of BGS strength in the human genome. D: Mean posterior estimates of α_W , as determined by separately fitting the model to alleles from each independent background selection strength bin. A linear model fit to the data (green line) supported statistically significant covariation between π/π_0 and α_W (p -value=0.0343). The black dashed line shows the predicted change in α_W as a function of B given the mean estimate of α_W .

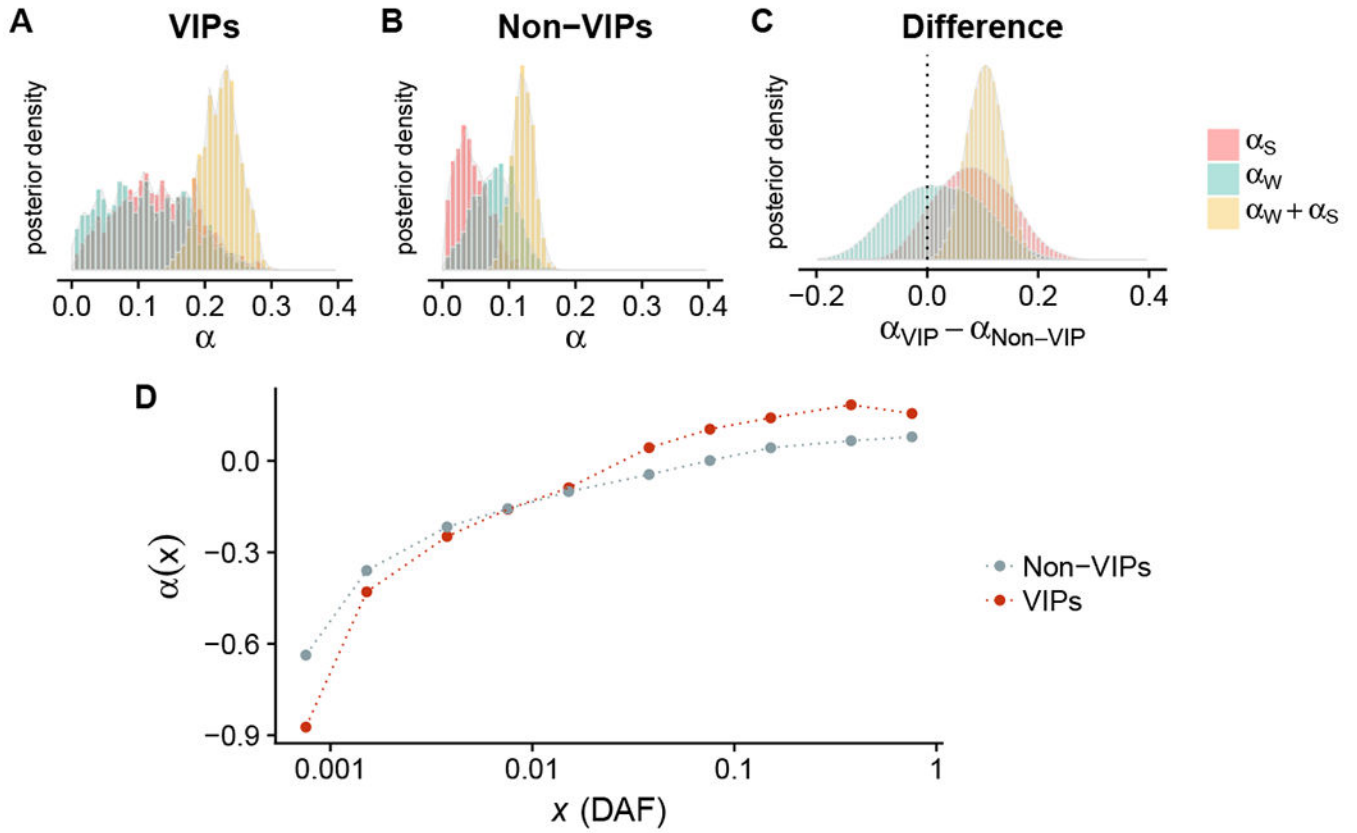


Figure 4: Virally-interacting genes support a high rate and strength of adaptation

A: Posterior distributions for α , α_W , and α_S for virus-interacting proteins (VIPs, 4,066 genes). B: The same quantities for non-VIPs (12,962 genes). C: The posterior distribution of the difference in α for VIPs and non-VIPs. D: $\alpha(x)$ as a function of derived allele frequency x for VIPs and non-VIPs as a function of derived allele frequency x , specifically at the values of x that we use for statistical inference.

Table 1:
Datasets and corresponding adaptation rates.

Estimated α values represent the mean of the posterior distribution. NS represents the number of nonsynonymous fixations and SYN represents the number of synonymous fixations. Values in parentheses represent the area of 95% highest posterior density. α : total adaptation rate; α_W : adaptation rate of weakly beneficial alleles; α_S : adaptation rate for strongly beneficial alleles.

Datasets & inferred adaptation rates					
Dataset	NS	SYN	$\hat{\alpha}$	$\hat{\alpha}_W$	$\hat{\alpha}_S$
Whole-egoism	29925	38135	0.135 (0.096,0.17)	0.097 (0.0,0.21)	0.041 (0.0,0.13)
VIPs	6249	10309	0.224 (0.17,0.28)	0.098 (0.0,0.24)	0.126 (0.018,0.26)
Non-VIPs	23676	27826	0.12 (0.09,0.15)	0.077 (0.01,0.13)	0.042 (0.0, 0.09)