Article Type: Letter

Application of a novel haplotype-based scan for local adaptation to study high-altitude adaptation in rhesus macaques

Zachary A. Szpiech[†]
szpiech@psu.edu

Department of Biology
Pennsylvania State University
University Park, PA 16801 USA

Institute for Computational and Data Sciences
Pennsylvania State University
University Park, PA 16801 USA

Department of Biological Sciences
Auburn University
Auburn, AL 36842, USA

*Taylor E. Novak**
tep0007@auburn.edu

Department of Biological Sciences
Auburn University
Auburn, AL 36842, USA

*Nick P. Bailey**
npb0015@auburn.edu

Department of Biological Sciences
Auburn University
Auburn, AL 36842, USA

Laurie S. Stevison[†]
lss0021@auburn.edu

Department of Biological Sciences
Auburn University
Auburn, AL 36842, USA

[†]Correspondence: szpiech@psu.edu, lss0021@auburn.edu
**Equal Contribution*

Running Title: ***High-altitude adaptation in rhesus macaques***

Keywords: Adaptation, genome-scan, high-altitude, EGLN1, selection, macaque

Abstract Word Count (300 max): 214
Total Word Count (5000 max): 5149

Abstract

When natural populations split and migrate to different environments, they may experience different selection pressures that can lead to local adaptation. To capture the genomic patterns of a local selective sweep, we develop XP-nSL, a genomic scan for local adaptation that compares haplotype patterns between two populations. We show that XP-nSL has power to detect ongoing and recently completed hard and soft sweeps, and we then apply this statistic to search for evidence of adaptation to high altitude in rhesus macaques. We analyze the whole genomes of 23 wild rhesus macaques captured at high altitude (mean altitude > 4000m above sea level) to 22 wild rhesus macaques captured at low altitude (mean altitude < 500m above sea level) and find evidence of local adaptation in the high-altitude population at or near 303 known genes and several unannotated regions. We find the strongest signal for adaptation at EGLN1, a classic target for convergent evolution in several species living in low oxygen environments. Furthermore, many of the 303 genes are involved in processes related to hypoxia, regulation of ROS, DNA damage repair, synaptic signaling, and metabolism. These results suggest that, beyond adapting via a beneficial mutation in one single gene, adaptation to high altitude in rhesus macaques is polygenic and spread across numerous important biological systems.

Impact Summary

When positive selection is ongoing or a beneficial mutation has recently fixed in a population, genetic diversity is reduced in the vicinity of the adaptive allele, and we expect to observe long homozygous haplotypes at high frequency. Here we develop a statistic that summarizes these expected patterns and compares between two

populations in order to search for evidence of adaptation that may have occurred in one but not the other. We implement this statistic in a popular and easy-to-use software package, and then apply it to study adaptation to high altitude in rhesus macaques.

Extreme environments pose a challenge to life on multiple fronts. Very high-altitude environments are one such example, with low atmospheric oxygen, increased ultraviolet light exposure, harsh temperatures, and reduced nutrition availability. In spite of these challenges, many plants and animals, including humans, have genetically adapted to cope with these hardships. Here we study two populations of rhesus macaques, one living at high altitude and one living close to sea level. We apply our novel statistic to compare their haplotype patterns between them to search for evidence of genetic changes that are indicative of adaptation to their environment.

We find evidence for adaptation at a critical gene that helps control physiological response to low-oxygen, one that has been the target of repeated convergent evolution across many species. We also find evidence for positive selection across a range of traits, including metabolic and neurological. This work helps to explain the evolutionary history of the rhesus macaque and furthers our understanding about the ways organisms genetically adapt to high-altitude environments.

Introduction

Selective sweeps produce regions of reduced genetic diversity in the vicinity of an adaptive mutation. These patterns manifest as long extended regions of homozygous haplotypes segregating at high frequency (Przeworski 2002; Sabeti *et al.* 2002; Kim and Nielsen 2004; Garud *et al.* 2015). In the event of a *de novo* mutation that is adaptive in a population, we expect the haplotype it resides on to rapidly rise in frequency in the population (called a 'hard' sweep). On the other hand, if an ancestrally segregating neutral or mildly deleterious allele turned out to be adaptive in a new environment, it would likely reside on two or more haplotypes, which would rapidly rise in frequency in the population (called a 'soft' sweep) (Hermisson and Pennings 2005; Pennings and Hermisson 2006). As both of these processes happen on a time scale faster than mutation or recombination can act to break up the sweeping haplotypes, we expect to observe long and low diversity haplotypes at high frequency in the vicinity of an adaptive mutation. However, if this mutation either does not exist or is not adaptive in a sister population, we would not expect a sweep to occur and thus we would not expect to observe similar haplotype patterns.

To capture these haplotype patterns and contrast them between a pair of populations, we develop XP-nSL, a haplotype-based statistic with good power to detect partial, fixed, and recently completed hard and soft sweeps by comparing a pair of populations. XP-nSL is an extension of nSL (Ferrer-Admetlla *et al.* 2014) and does not require a genetic recombination map for computation. The lack of dependence on a recombination map is important, as other statistics for identifying positive selection are biased towards low-recombination regions (O'reilly *et al.* 2008), but the approach taken by nSL has been shown to be more robust (Ferrer-Admetlla *et al.* 2014). Both nSL and XP-nSL summarize haplotype diversity by computing the mean number of sites in a region that are identical-by-state across all pairs of haplotypes. Whereas nSL contrasts between haplotype sets carrying an ancestral or a derived allele in a single population, XP-nSL contrasts between haplotype sets in two different populations, allowing it to detect local adaptation.

An extreme example of adaptation to a local environment is the transition to high-altitude living. Organisms living at high altitude are confronted with many challenges, including a low-oxygen atmosphere and increased ultraviolet light exposure, and these harsh environments inevitably exert strong selection pressure. Indeed, adaptation to high-altitude living has been studied extensively across many organisms from plants, including monocots (Gonzalo-Turpin and Hazard 2009; Ahmad *et al.* 2016) and dicots (Kim and Donohue 2013; Liu *et al.* 2014; Munne-Bosch *et al.* 2016; Guo *et al.* 2018), to numerous animals including amphibians (Yang *et al.* 2016), canids (Li *et al.* 2014; Wang *et al.* 2014; Wang *et al.* 2020), humans (Bigham *et al.* 2009; Bigham *et al.* 2010; Xu *et al.* 2010; Yi *et al.* 2010; Peng *et al.* 2011; Huerta-Sanchez *et al.* 2013; Huerta-Sanchez *et al.* 2014; Jeong *et al.* 2014), yaks (Qiu *et al.* 2012), birds (Cai *et al.* 2013; Qu *et al.* 2013; Wang *et al.* 2015; Graham and Mccracken 2019), boars (Li *et al.* 2013), mice (Storz *et al.* 2007; Cheviron *et al.* 2012; Schweizer *et al.* 2019; Storz *et al.* 2019; Velotta *et al.* 2020), moles (Campbell *et al.* 2010), antelope (Ge *et al.* 2013), and horses (Hendrickson 2013). Liu *et al.* (2018) recently sequenced and published the whole genomes of 79 wild-born Chinese rhesus macaques collected from multiple sites in China. Among these animals, 23 were sampled from far western Sichuan province in a region with mean altitude > 4000 m above sea level (Liu *et al.* 2018), providing an opportunity to study the genetics of local adaption to high altitude in rhesus macaques.

Rhesus macaques are the second most widely distributed primate, with a range extending from Afghanistan to Vietnam and from a latitude of 15 to 38 degrees north (Fooden 2000). Early ancestors of the macaque migrated out of Africa to the Eurasian continent approximately 7 mya—the earliest catarrhine fossils on the continent are macaque-like (Stewart and Disotell 1998). Modern rhesus macaques trace a recent origin to Southeast Asia, with a major migratory split occurring approximately 162 kya separating the ancestors of modern Indian and Chinese rhesus macaques (Hernandez *et al.* 2007). Macaques have proven to be quite evolutionarily successful, demonstrating ecological flexibility and adaptability via developmental plasticity and behavioral changes (Richard *et al.* 1989; Madrid *et al.* 2018). Other studies have looked at how the rhesus macaque radiation has led to population-level adaptation to climate and food availability (Liu *et al.* 2018).

Here we test and evaluate our XP-nSL statistic and apply it to study the genomic consequences of high-altitude living in the rhesus macaque. We use it to compare the haplotype patterns of the 23 animals from the high-altitude population with another 22 that Liu *et al.* (2018) sampled in lower-lying regions in eastern China with a mean altitude < 500 m above sea level.

Methods

Data Preparation

Liu *et al.* (2018) generated whole genome sequencing data for 79 Chinese macaques and called all biallelic polymorphic sites according to GATK best practices using rheMac8, identifying 52,534,348 passing polymorphic autosomal sites. We then filter all loci with > 10% missing data leaving 35,639,395 biallelic sites. Next, the program SHAPEIT v4.1.2 (Delaneau *et al.* 2019) was used to phase haplotypes in the full data set with a genetic map that was available for rheMac8 (Bcm-Hgsc 2020). SHAPEIT performs imputation during phasing for any missing genotypes. Our analyses here focus on 45 of the 79 samples from Liu *et al.* (2018), representing 23 from high-altitude regions of China (*M. m. lasiotis*) and 22 from low-altitude regions of China (*M. m. littoralis*), based on capture location information. See Table S1 for individual IDs used.

Liu *et al.* (2018) also inferred joint demographic histories for their five populations, and we extract the demographic parameters for our two of interest. This demographic history is recapitulated in Fig. 1 with detailed parameters given in Table 1, which are then used for simulations to test XP-nSL.

A Statistic for Detecting Local Adaptation

We developed a cross-population haplotype-based statistic, XP-nSL, to scan for regions of the genome implicated in local adaptation between two populations by extending nSL (Ferrer-Admetlla *et al.* 2014), each of which is defined below.

Consider the sets $A(k)$ and $D(k)$, representing the set of haplotypes at site k carrying the ancestral or derived allele, respectively, and let $n_A(k) = |A(k)|$ and $n_D(k) = |D(k)|$. Next $L_{ij}(k)$ is defined as the number of consecutive sites at which haplotype i and j are identical-by-state (IBS) in the interval containing site k. Then nSL at site k is $nS_L(k) = \log\frac{SL_A(k)}{SL_D(k)}$, where $SL_A(k) = \binom{n_A(k)}{2}^{-1} \sum_{i<j\in A(k)} L_{ij}(k)$ and $SL_D(k) = \binom{n_D(k)}{2}^{-1} \sum_{i<j\in D(k)} L_{ij}(k)$. $SL_A(k)$ and $SL_D(k)$ represent the mean $L_{ij}(k)$ over all pairs of haplotypes carrying either the ancestral or derived allele at locus k, respectively. nSL scores are then normalized genome-wide in site-frequency bins either with respect to the empirical background or neutral simulations with a matching demographic history. The nSL computation is illustrated in Ferrer-Admetlla *et al.* (2014). nSL is implemented in `nsl` (Ferrer-Admetlla *et al.* 2014) and `selscan` v1.1.0+ (Szpiech and Hernandez 2014).

XP-nSL is defined similarly, except instead of comparing sets of haplotypes containing ancestral or derived alleles, it compares sets of haplotypes between two different populations. Let $P_1(k)$ and $P_2(k)$, represent the set of haplotypes at site k in population 1 and population 2, respectively, and let $n_{P_1}(k) = |P_1(k)|$ and $n_{P_2}(k) = |P_2(k)|$. Then XP-nSL at site k is $XPnS_L(k) = \log\frac{SL_{P_1}(k)}{SL_{P_2}(k)}$, where $SL_{P_1}(k) = \binom{n_{P_1}(k)}{2}^{-1} \sum_{i<j\in P_1(k)} L_{ij}(k)$ and $SL_{P_2}(k) = \binom{n_{P_2}(k)}{2}^{-1} \sum_{i<j\in P_2(k)} L_{ij}(k)$. $SL_{P_1}(k)$ and $SL_{P_2}(k)$ represent the mean $L_{ij}(k)$ over all pairs of haplotypes in population 1 or population 2 at locus k, respectively. XP-nSL scores are then normalized genome-wide either with respect to the empirical background or neutral simulations with a matching demographic history. The XP-nSL computation is illustrated in Fig. 2 with a toy example. We implement XP-nSL in `selscan` v1.3.0+ (Szpiech and Hernandez 2014) to facilitate wide adoption. It is worth noting that XP-nSL is analogous to XP-EHH (Sabeti *et al.* 2007) as nSL is analogous to iHS (Voight *et al.* 2006).

The goal of these statistics is to capture a signal of extended regions of low diversity on sweeping haplotypes (emblematic of an ongoing or recently completed selective sweep) within a population (nSL) or on sweeping haplotypes in one population versus another (XP-nSL). When XP-nSL scores are positive this suggests evidence for a hard or soft sweep in population 1, and when XP-nSL scores are negative this suggests evidence for a hard or soft sweep in population 2.

Simulations

In order to test the ability of XP-nSL to detect ongoing and recently completed hard and soft sweeps, coalescent simulations were performed conditional on an allele frequency trajectory with the program `discoal` (Kern and Schrider 2016). `discoal` simulates an allele frequency trajectory for a single non-neutral allele backwards in time and then simulates a neutral coalescent process conditional on this trajectory. This takes advantage of the speed and efficiency of the coalescent while still being able simulate genetic diversity patterns in the vicinity of a non-neutral locus.

All simulations were run with a two-population divergence demographic history (Fig. 1 and Table 1), as inferred by Liu *et al.* (2018), and given by the following `discoal` command line arguments `-p 2 46 44 -en 0 1 0.230410476572876 -en 0.071964666275442 1 0.278345739259351 -en 0.086558359329153 1 1.282885999320505 -ed 0.146214905642895 1 0 -en 0.146214905642895 0 4.089940389782871`. Here 46 haplotypes were sampled from population 0, which has the demographic history of the high-altitude population, and 44 haplotypes were sampled from population 1, which has the demographic history of the low-altitude population. A mutation rate of $\mu = 2.5 \times 10^{-8}$ (Fan *et al.* 2018) was used, along with a recombination rate of $r = 5.126 \times 10^{-9}$, which was computed as the genome-wide mean rate from the rheMac8 recombination map (Bcm-Hgsc 2020). A 500 kb region was simulated, thus giving a scaled mutation rate and scaled recombination parameters for `discoal` as `-t 809.425 -r 165.967`. 5,349 replicates of neutral sequence were simulated under this model, representing approximately the entire macaque genome minus 500 kb. Thus, the total simulated length of all neutral regions plus one selected region is approximately equal to the macaque genome length.

For non-neutral simulations, sweep scenarios were simulated with a positive additive selection coefficient $s \in \{0.01, 0.02, 0.05\}$, which is provided to `discoal` as a scaled selection coefficient $2N_1 s \in \{323.77, 647.54, 1618.85\}$ (`discoal` flag `-a`). Soft sweeps are simulated as a mutation that arose neutral and turned beneficial at a particular establishment frequency $e \in \{0.01,0.02,0.03,0.04,0.05,0.10,0.20\}$ (`discoal` flag `-f`). Hard sweeps are simulated from a *de novo* mutation that was never neutral, i.e. $e = 0$. Finally, sweeps were conditioned on having either reached a certain frequency in the population at the time of sampling, $f \in \{0.7, 0.8, 0.9, 1.0\}$ (`discoal` flag `-c`), or that the adaptive mutation reached fixation some number of generations prior to sampling, $g \in \{50, 100, 200\}$, which is provided to `discoal` in coalescent units $g/2N_1 \in \{1.5443 \times 10^{-3}, 3.0886 \times 10^{-3}, 6.1774 \times 10^{-3}\}$ (`discoal` flag `-ws`). For the sake of being conservative in our estimation of power (Voight *et al.* 2006; Sabeti *et al.* 2007), it was assumed the actual adaptive mutation remains unsampled (`discoal` flag `-h`) even though whole genome sequencing data are being analyzed. For each combination of parameter values, 500 replicates were simulated.

Neutral Simulations of Mismatched Histories

We also generate neutral simulations for three mismatched demographic histories, in order to study how a mismatched demographic history may influence the power and false positive rates of XP-nSL if used as a normalization baseline. We name these mismatched histories “Rand”, “Under”, and “Over” and generate 5,349 replicates for each one.

For the “Rand” history, each parameter is (uniformly) randomly chosen from within the 95% CI as inferred by (Liu *et al.* 2018) (see Table 1). The parameters are resampled for each replicate of the “Rand” history, and the condition $T_1 < T_2 < T_3$ is enforced. For the “Under” history, the present-day population size parameters N_1 and N_2 are the only ones modified. They are set to $N_1 = 8255$ at the extreme low end of the 95% CI and $N_2 = 5006$ at the extreme high end of the 95% CI. This represents a scenario where the difference in population sizes is underestimated. For the “Over” history, once again the present-day population size parameters N1 and N2 are the only ones modified. They are set to $N_1 = 26825$ at the extreme high end of the 95% CI and $N_2 = 660$ at the extreme low end of the 95% CI. This represents a scenario where the difference in population sizes is overestimated.

Detecting Local Adaptation in Real Data

From the phased data set, animals captured at high altitude (n = 23) and animals captured at low altitude (n = 22) were subset (see Table S1). Using `selscan` v1.3.0 (Szpiech and Hernandez 2014) to compute raw XP-nSL scores across the genome (`selscan` flags –xpnsl –vcf high-altitude.vcf –vcf-ref low-altitude.vcf), scores were then normalized using the genome-wide empirical background with `selscan`’s `norm` v1.3.0 program (`norm` flag --xpnsl). The low-altitude population was used as the reference population, so positive XP-nSL scores correspond to long homozygous haplotypes and a possible sweep in the high-altitude population compared to the low-altitude population, and vice versa for negative XP-nSL scores. A Manhattan plot of genome-wide normalized XP-nSL scores > 2 is plotted in Fig. S7.

In order to identify regions implicated as potentially adaptive, we search for clusters of extreme scores along a chromosome. Using `selscan`’s companion program `norm v1.3.0`, the genome is divided into non-overlapping 100kb regions and both the maximum XP-nSL score and the fraction of XP-nSL scores > 2 are computed (`norm` flags –xpnsl –bp-win –winsize 100000). `norm` then creates 10 quantile bins (--qbins 10) for windows with more than 10 sites per window (--min-snps 10) and identifies the top

1% of windows with the highest fraction of extreme scores (Fig. S6 and Table S4). Each window is then annotated with the ensembl rheMac8 gene list, and a maximum XP-nSL score is assigned to a given gene based on the max-score in the 100 kb window with which it overlaps. If a gene overlaps more than one 100 kb window, it is assigned the top max-score from among the windows.

Results

Power Analysis of XP-nSL

First, we evaluate the performance of XP-nSL based on simulations. After computing XP-nSL for all sites in all simulations, scores were normalized by subtracting the mean and dividing by standard deviation of the neutral simulations, giving the neutral scores an approximately N(0,1) distribution (Fig. S1).

We consider the maximum score in a 100 kb interval as way to identify regions under positive selection similar to Voight *et al.* (2006). To get the null distribution of max-scores, the maximum score is computed in the central 100 kb of each neutral simulation. The distribution of max-scores in neutral simulations had a median of 2.093 with 95% of the mass between 0.804 and 3.492, which is represented in Fig. 3 as a solid horizontal black line (median) and two dashed horizontal black lines (95% interval). Next, the maximum score is computed in the central 100 kb of all non-neutral simulations, and the median and 95% intervals are plotted for each parameter combination. Fig. 3 shows good separation between the neutral distribution of max-scores and the distribution of max-scores for a range of non-neutral parameters, suggesting that our statistic can distinguish between neutral and non-neutral scenarios. Note that soft sweeps that start at 0.1 or 0.2 frequency see the least separation from the neutral distribution. To evaluate the power of the max-score statistic, the 99th percentile of the max-score distribution is computed in the neutral distribution ($neutral_{99} = 3.792$) and power is calculated as the mass of each non-neutral max score distribution $> neutral_{99}$. With this approach, even if the entire genome is neutrally evolving, at most 1% of the genome will be identified as putatively under selection thus fixing the false positive rate at 1% at most. Results are plotted in Fig. 4A, which shows good power to detect both incomplete and completed sweeps. For soft sweeps that start at a frequency ≤ 0.03, power is $> 75\%$ when the sweep is near or just past fixation; the ability to detect soft sweeps falls off for sweeps that start > 0.03 frequency.

Next, we consider that due to linkage disequilibrium consecutive scores will be correlated, and we should therefore expect clusters of extreme scores in true non-neutral regions. We thus consider a window-based approach to identify selected regions similar to Voight *et al.* (2006), where the top 1% windows with a high number of extreme scores are identified. Taking the central 100 kb region of each simulation, the fraction of XP-nSL scores > 2 (representing approximately the highest 2% of all neutral scores) is computed. Since each 100 kb window has a variable number of sites within it, windows with fewer sites are more likely to have a higher fraction of extreme scores by chance. Thus, windows are binned by number of sites into 10 quantile bins, and the top 1% of windows with the highest fraction of extreme scores in each bin establishes the threshold beyond which a window is taken as putatively selected, as in Voight *et al.* (2006). With this approach, even if the entire genome is neutrally evolving, at most 1% of the genome will be identified as putatively under selection thus fixing the false positive rate at 1% at most. Power is computed for each non-neutral parameter set as the proportion of replicates for which the central 100 kb exceeds the 1% threshold as calculated from neutral simulations. The results are plotted in Fig. 4B, which shows improved power over the max-score approach across a wider range of parameters. Indeed, using the window-based method, power to detect soft sweeps improves substantially across the parameter space, with $> 75\%$ power to detect soft sweeps at or near fixation that started at frequency ≤ 0.05.

We next consider how XP-nSL power compares to nSL, XP-EHH, and F_{ST}. We compare to nSL since XP-nSL is an extension of it, to XP-EHH as it is a similar haplotype-based two-population selection statistic, and to F_{ST} as it a popular two-population method used to infer local adaptation. For all simulations, nSL and XP-EHH are computed using `selscan` v1.3.0 (Szpiech and Hernandez 2014). Normalization, identification of top windows, and power calculation was done as described above for XP-nSL. F_{ST} is computed using `VCFtools` v0.1.16 (Danecek *et al.* 2011), which implements Weir and Cockerham's formulation (Weir and Cockerham 1984), in 100kb windows for all simulations. The 99th percentile F_{ST} value was determined from the central 100kb window among all neutral simulations. Power was then computed for each non-neutral parameter set as the proportion of replicates for which the F_{ST} value of the central 100kb window is greater than the neutral threshold as calculated from neutral simulations.

Fig. 5 shows the difference in power between XP-nSL and each statistic (raw power for nSL, XP-EHH, and F_{ST} shown in Fig. S4) over the simulated parameter space, where positive values indicate that XP-nSL has more power than the comparison statistic. XP-nSL is compared to nSL in Fig. 5A, which shows XP-nSL improves on nSL across nearly the entire parameter space and especially for sweeps that have fixed in the recent past. XP-nSL is compared to XP-EHH in Fig. 5B, which shows XP-nSL improving on XP-EHH for soft sweeps ($e > 0$). Finally, XP-nSL is compared to F_{ST} in Fig. 5C, which shows XP-nSL improving on F_{ST} mostly for incomplete sweeps, although F_{ST} performed better post-fixation for certain soft sweep scenarios ($e \geq 0.05$).

Caveats for Using Neutral Simulations to Normalize Real Data

In principle, one could use matched neutral simulations as a normalization baseline when analyzing real data. However, we can only recommend this approach when the populations being studied have very well characterized (1) joint demographic histories, (2) mutation rates, (3) and recombination rates, as a mismatch can skew the power and false positive rates of the statistic. To illustrate this point, we generated three mismatched sets of neutral simulations “Rand”, “Under”, and “Over” (see Methods) to use as a normalization baseline for our original simulations.

When neutral simulations are normalized with the correct demographic history, they approximately follow a standard normal distribution (Fig. S1 and Fig. S2), however the distribution of neutral scores gets badly distorted when one of the mismatched histories is used (Fig. S2). These distortions have practical consequences for making inferences. Power was calculated for XP-nSL using the window-based method described above but using each mismatched history as a normalization baseline (Fig. S3). The false positive rate for each scenario was also estimated by calculating the proportion of neutral simulations that are identified as under selection for each scenario (Table S2). Fig. S3A-B shows that for the “Rand” and “Under” normalization scenarios, power was greatly reduced across the whole parameter space. Only hard sweeps with the strongest selection coefficients post-fixation were likely to be identified. False positive rates for these scenarios were at 0 (Table S3), whereas for a matched history it was estimated at 9.908×10^{-3}. Fig.S3C shows that, for the "Over" normalization scenario, power was uniformly excellent, however when analyzing neutral simulations, the false positive rate was estimated at 0.9538. This suggests that using such a badly matched demographic history for normalization creates a near complete inability to distinguish between neutral and selected regions.

Identifying Evidence for Adaptation to High Altitude in Rhesus Macaques

Next, we analyzed the pair of rhesus macaque populations using XP-nSL, searching for evidence of local adaptation in the high-altitude population. Using the low-altitude population as the reference population, normalized $XPnS_L > 0$ corresponds to longer and higher frequency haplotypes in the high-altitude population, with very large positive scores and clusters of large scores suggesting evidence for positive selection. Dividing the genome into 100 kb windows, the maximum XP-nSL score of that region is assigned to each gene in it (see Methods), thus multiple genes may have the same max-score by virtue of being in the same 100kb window. Genes overlapping multiple 100kb windows were assigned the top max-score among the windows.

Using the per-gene max-scores, PANTHER (Mi *et al.* 2019) gene ontology categories were tested for enrichment of high scores, where significance suggests an enrichment of signals of positive selection among genes involved. Significant terms related to regulation of ion transport and synaptic signaling (Table 2), each of which are affected by hypoxic conditions (Karle *et al.* 2004; Corcoran and O'connor 2013).

From the genomic regions that were identified to contain a high proportion of extreme positive scores (see Methods), 303 annotated genes were found across 270 regions. A permutation test (10,000 replicates) that randomly shuffles 270 100kb regions around the genome indicates that this is substantially fewer than one would expect by chance ($p = 1.4 \times 10^{-3}$; Fig. S5), indicating that the method is not simply randomly picking gene regions. These regions, their characteristics, and the genes contained therein are given in Table S2. A PANTHER (Mi *et al.* 2019) gene ontology overrepresentation test indicates a 9.04-fold enrichment of genes associated with monooxygenase activity ($FDR = 4.47 \times 10^{-2}$).

The monooxygenases in the selected regions include FMO2, FMO5, CYP2C8, CYP2C9, CYP2C93, and ENSMMUT00000011129. These genes are important for the metabolism of oxygen and the generation of reactive oxygen species (ROS) (Krueger and Williams 2005). Under the physiological stress of a low-oxygen environment, ROS levels increase and cause oxidative damage, and, in humans, long-term adaptation to high altitudes includes adaptation to oxidative damage (Janocha *et al.* 2017). Indeed, AOX1 is also identified in our top regions, mutations in which have been shown to affect ROS levels in humans (Foti *et al.* 2017).

The genome-wide top ten 100 kb windows based on the percentage of extreme XP-nSL scores are summarized in Table 3, and these windows overlap several genes, including EGLN1. The EGLN1 locus is directly adjacent to the single strongest selection signal identified in the entire genome (Fig. 6). This region has the third highest cluster of extreme scores (Table 3), contains the highest XP-nSL score in the entire genome (chr1:207,698,003, $XPnS_L = 6.54809$), and contains six of the top ten genome wide XP-nSL scores (colored dark red in Fig. 6). EGLN1 is a regulator of oxygen homeostasis (To and Huang 2005) and is a classic target for adaptation to low-oxygen levels, having repeatedly been the target of adaptation in numerous organisms living at high altitude around the world (Bigham *et al.* 2009; Bigham *et al.* 2010; Jeong *et al.* 2014; Graham and Mccracken 2019). In addition to EGLN1, other genes related to lung function, oxygen use, and angiogenesis had evidence for local adaptation between low- and high-altitude populations: TRPM7, RBPJ, and ENSMMUT00000040566 (Table S2). TRPM7 downregulation in a hypoxia-induced rat model was associated with pulmonary hypertension (PAH) (Xing *et al.* 2019). ENSMMUT00000040566 is a MAPK6 ortholog, which interacts with EGLN3 (Rodriguez *et al.* 2016), and both it and RBPJ are involved in angiogenesis (Ramasamy *et al.* 2014).

Due to the reduced oxygen levels at high altitudes, we expect genes involved in metabolism and respiration may be under positive selection. Indeed, MDH1 encodes a critical enzyme in the citrate cycle (Tanaka *et al.* 1996) and is found in the top ten genome-wide regions (Table 3). A paralog of MDH1, MDH1B, has been previously identified as a target of selection in humans living at high altitude (Yi *et al.* 2010). ACADM and COX15 are also found in putatively adaptive regions (Table S2). Mutations in and differing expression levels of ACADM are related to oxidative stress and mitochondrial dysfunction in human disease (Xu *et al.* 2018). COX15 is involved in oxidative phosphorylation (Alston *et al.* 2017), and cytochrome c oxidase (COX) genes have previously been identified as under selection in primates relative to other mammals (Osada and Akashi 2012).

High-altitude environments present a particular metabolic challenge to organisms that must maintain a stable internal body temperature (Rosenmann and Morrison 1974; Hayes and Chappell 1986; Chappell and Hammond 2004), a result of increased oxygen demand from aerobic thermogenesis conflicting with lower oxygen availability. It has been shown that highland deer mice have adapted by increased capacity to metabolize lipids compared to lowland deer mice (Cheviron *et al.* 2012), and previous studies in rhesus macaques have shown there may be drastic differences in diets between high- and low-altitude populations (Zhao 2018). Indeed, in the high-altitude macaque population studied here, genes related to lipid and fat metabolism (DOCK7, ST6GALNAC5, ANGPTL3, and ACACA) were found in putative adaptive regions. Across human populations, these genes are all responsible for varying blood levels of fatty acids (Guo *et al.* 2016; Dewey *et al.* 2017; Hebbar *et al.* 2018). Furthermore, ACACA has been shown to vary fatty acid blood concentrations and be differentially expressed in highland versus lowland swine populations (Shang *et al.* 2019).

STXBP5L also appears in the top ten regions (Table 3) and is involved in vesicular trafficking and neurotransmitter release (Kumar *et al.* 2015). As primate brains use large amounts of oxygen and energy to function (Osada and Akashi 2012) signatures of selection on neurological genes may be expected across populations living at altitudes with differing oxygen levels. In addition to STXBP5L, several genes related to neural development and synaptic formation (JAG2, TRPM7, DOCK7, NSG2, AUTS2) were identified (Table S2). JAG2 is involved in the Notch signaling pathway. While Notch signaling is involved in many developmental and homeostatic processes, its role in neuronal differentiation in the mammalian brain is notable in this context

(Cardenas *et al.* 2018). TRPM7, in addition to its association with PAH, plays a role in hypoxic neuronal cell death (Aarts *et al.* 2003).

DNA damage, including double strand breaks and pyrimidine dimerization, can manifest as a result of oxidative stress (Ye *et al.* 2016) or increased exposure to UV radiation (Zhang *et al.* 2000; Greinert *et al.* 2012), both of which increase at high altitudes. Three DNA damage repair genes are in our set of genes identified in positively selected regions. The ring finger protein RNF138 is in our list of top ten genomic windows (Table 3) and has been shown to promote DNA double-strand break repair (Ismail *et al.* 2015). Furthermore, the DNA polymerase POLH also appears in a putatively adaptive region (Table S2) and is known to be able to efficiently bypass pyrimidine dimer lesions (Zhang *et al.* 2000). Finally, PAXIP1 appears in Table S2 as well and has been shown to promote repair of double strand breaks through homologous recombination (Wang *et al.* 2010).

Interestingly, within the putatively selected region that includes PAXIP1 is an uncharacterized long non-coding RNA (lncRNA), ENSMMUT00000081951, that was recently annotated in the rheMac10 genome build. This lncRNA has high sequence similarity to a lncRNA on the same synteny block in humans called PAXIP1-AS1. In human pulmonary artery smooth muscle cells, knockdown of PAXIP1-AS1 leads to an abnormal response to PAH where migration and proliferation of cells is reduced, and overexpression of PAXIP-AS1 leads to apoptosis resistance (Jandl *et al.* 2019). Although, we note that any link between PAH and ENSMMUT00000081951 in rhesus macaques is highly speculative at this point.

Discussion

When populations split and migrate, they may adapt in different ways in response to their local environments. Genetic adaptations that arise and sweep through the population leave a characteristic genomic pattern of long haplotypes of low diversity and high frequency. We develop a two-population haplotype-based statistic, XP-nSL, to capture these patterns. With good power to detect hard and soft sweeps that occur in one population but not another, XP-nSL can identify positively selected regions of the genome likely the result of local adaptation. We apply this statistic to genomes sampled from a pair of wild-born populations of Chinese rhesus macaques, inferred to have diverged approximately 9500 generations ago, one of which lives at high altitude in far western Sichuan province, and the other that lives close to sea level.

Life at high altitude presents extreme biological challenges, including low atmospheric oxygen, increased UV exposure, harsh winters, and reduced nutrition availability, which create strong selection pressure. Organisms that survive and persist are likely to be carrying genetic mutations that confer an advantage for living in such harsh environments. Common targets for adaptation to such an environment include genes related to hypoxia, regulation of ROS, DNA damage repair, and metabolism (Cheviron and Brumfield 2012; Witt and Huerta-Sanchez 2019; Storz and Cheviron 2021). Indeed, in the high-altitude macaque population, we identify a strong signal of positive selection at the EGLN1 locus (Fig. 6), a classic target for adaptation to low-oxygen environments, in addition to 302 other genes, many of which are related to the myriad environmental selection pressures expected in high-altitude environments. As has been suggested previously for other organisms (Cheviron and Brumfield 2012; Bigham and Lee 2014; Simonson 2015; Witt and Huerta-Sanchez 2019; Storz and

Cheviron 2021), these results suggest that, rather than a single adaptive mutation at a single locus, adaptation to this extreme environment by rhesus macaques is polygenic and acts through multiple biological systems.

Acknowledgments

The authors would like to thank members of the Stevison Lab for helpful discussions, Lawrence Uricchio for helpful comments on early versions of the manuscript, and two very helpful anonymous reviewers. This work was supported by start-up funds from the Department of Biological Sciences at Auburn University (LSS) and the Department of Biology at the Pennsylvania State University (ZAS). ZAS was partially supported by NSF-DEB EAGER No. 1939090 (LSS). Portions of this research were performed on the Pennsylvania State University's Institute for Computational Data Sciences' Roar supercomputer.

Author Contributions

ZAS and LSS conceived of the study. ZAS performed all simulations and genomic analyses and implemented novel statistics. ZAS, TEN, and NPB characterized gene functions and ontologies with contributions from LSS. ZAS wrote the manuscript with contributions from NPB, TEN, and LSS. All authors read and approved of the manuscript.

Data Accessibility

Macaque whole genome VCFs are available at http://dx.doi.org/10.5524/100484. Selection scan data available at https://doi.org/10.5061/dryad.kkwh70s40.

Tables

Table 1. Demographic parameters used for simulations with 95% confidence intervals from (Liu *et al.* 2018). T values are given in number of generations before present. N values represent diploid effective population size.

Parameter	Value	95% CI
T_1	4660	(651, 5034)
T_2	5605	(3962, 10468)
T_3	9468	(4563, 14047)
N_1	16188	(8255, 26825)
N_2	3730	(660, 5006)
N_3	4506	(572, 124135)
N_4	20768	(1710, 88242)
N_5	66210	(20581, 301430)

Table 2. Gene ontology enrichment analysis results based on maximum XP-nSL scores per gene. Significant GO terms are enriched for high XP-nSL scores.

PANTHER GO-Slim Biological Process	Gene Ontology ID	p-value	FDR
anterograde trans-synaptic signaling	GO:0098916	4.43×10^{-5}	1.52×10^{-2}
chemical synaptic transmission	GO:0007268	4.43×10^{-5}	1.83×10^{-2}
regulation of transport	GO:0051049	3.60×10^{-5}	1.85×10^{-2}
trans-synaptic signaling	GO:0099537	6.32×10^{-5}	1.86×10^{-2}
synaptic signaling	GO:0099536	1.03×10^{-4}	2.66×10^{-2}
ion transport	GO:0006811	1.33×10^{-4}	2.75×10^{-2}
transmembrane transport	GO:0055085	1.57×10^{-4}	2.94×10^{-2}
regulation of ion transport	GO:0043269	1.30×10^{-4}	2.97×10^{-2}
metal ion transport	GO:0030001	2.21×10^{-4}	3.04×10^{-2}
regulation of localization	GO:0032879	2.13×10^{-4}	3.14×10^{-2}
regulation of ion transmembrane transport	GO:0034765	2.66×10^{-4}	3.23×10^{-2}
inorganic ion transmembrane transport	GO:0098660	3.01×10^{-4}	3.45×10^{-2}
sodium ion transport	GO:0006814	3.36×10^{-4}	3.65×10^{-2}
ion transmembrane transport	GO:0034220	4.50×10^{-4}	4.42×10^{-2}
regulation of transmembrane transport	GO:0034762	4.40×10^{-4}	4.54×10^{-2}
cation transport	GO:0006812	4.87×10^{-4}	4.57×10^{-2}

Table 3. The top ten 100 kb genomic regions as ranked by percentage of scores greater than 2. Concatenated region represents the genomic region merged with adjacent top 1% regions. Max score represents the max XP-nSL score in the concatenated region. Genes gives all genes overlapping the concatenated region.

Genomic Region	% > 2	Flanking Regions	Concatenated Region	Max Score	Genes
*chr2:42300001-42400000	81.62%	3	chr2:42200001-42600000	5.42043	STXBP5L
chr3:99100001-99200000	78.18%	3	chr3:99100001-99500000	5.54282	-
chr1:207600001-207700000	***77.39%***	***1***	***chr1:207600001-207800000***	***6.54809***†	***EGLN1, TSNAX***
chr10:69700001-69800000	76.60%	4	chr10:69700001-70200000	4.89388	PITPNB, ENSMMUT00000079195.1, TTC28
chr13:64300001-64400000	74.01%	1	chr13:64300001-64500000	4.17728	WDPCP, MDH1, ENSMMUT00000070474.1
chr7:96400001-96500000	73.86%	1	chr7:96400001-96600000	4.58886	FAM177A1, PPP2R3C, ENSMMUT00000039911.2, ENSMMUT00000027928.3
chr10:70400001-70500000	70.24%	0	chr10:70400001-70500000	3.69192	TTC28
*chr2:42400001-42500000	65.86%	3	chr2:42200001-42600000	5.42043	STXBP5L
chr7:120400001-120500000	62.54%	0	chr7:120400001-120500000	5.86114	KIAA0586, ENSMMUT00000059161.1
chr18:20600001-20700000	62.42%	0	chr18:20600001-20700000	5.37386	RNF138

***These regions are contained in the same concatenated region.**
†Top genome-wide score. This region contains 6 of the top 10 genome-wide scores.

Table S1. Population classification by individual ID.
See TableS1.xlsx

Table S2. List of genomic windows with the top 1% highest fraction of extreme XP-nSL scores.
See TableS2.xlsx

Table S3. Estimates of false positive rates for various demographic history normalization scenarios.

Demographic History	Estimated False Positive Rate
Matched	9.908×10^{-3}
“Rand”	0
“Under”	0
“Over”	0.9538

Table S4. Bin boundaries (# of sites), number of windows per bin, and top 1% thresholds for the XP-nSL analysis of rhesus macaques.

Bin Boundaries (min # sites, max # sites)	# Windows in Bin	Top 1% Threshold (Fraction of Scores > 2)
(11, 731)	2672	0.346755
(732, 987)	2680	0.262945
(988, 1155)	2676	0.207089
(1156, 1285)	2675	0.196678
(1286, 1390)	2663	0.226605
(1391, 1484)	2687	0.19274
(1485, 1576)	2665	0.163597
(1577, 1681)	2648	0.169275
(1682, 1824)	2666	0.164025
(1825, 5766)	2666	0.153176

Figures

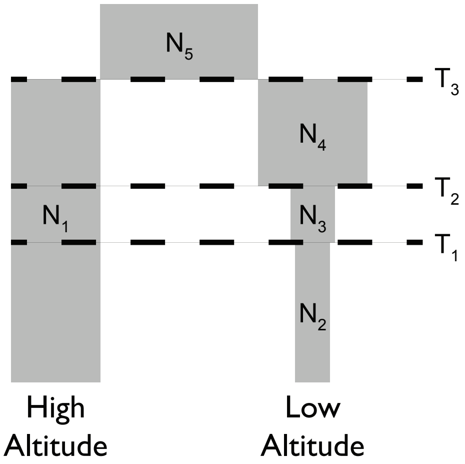
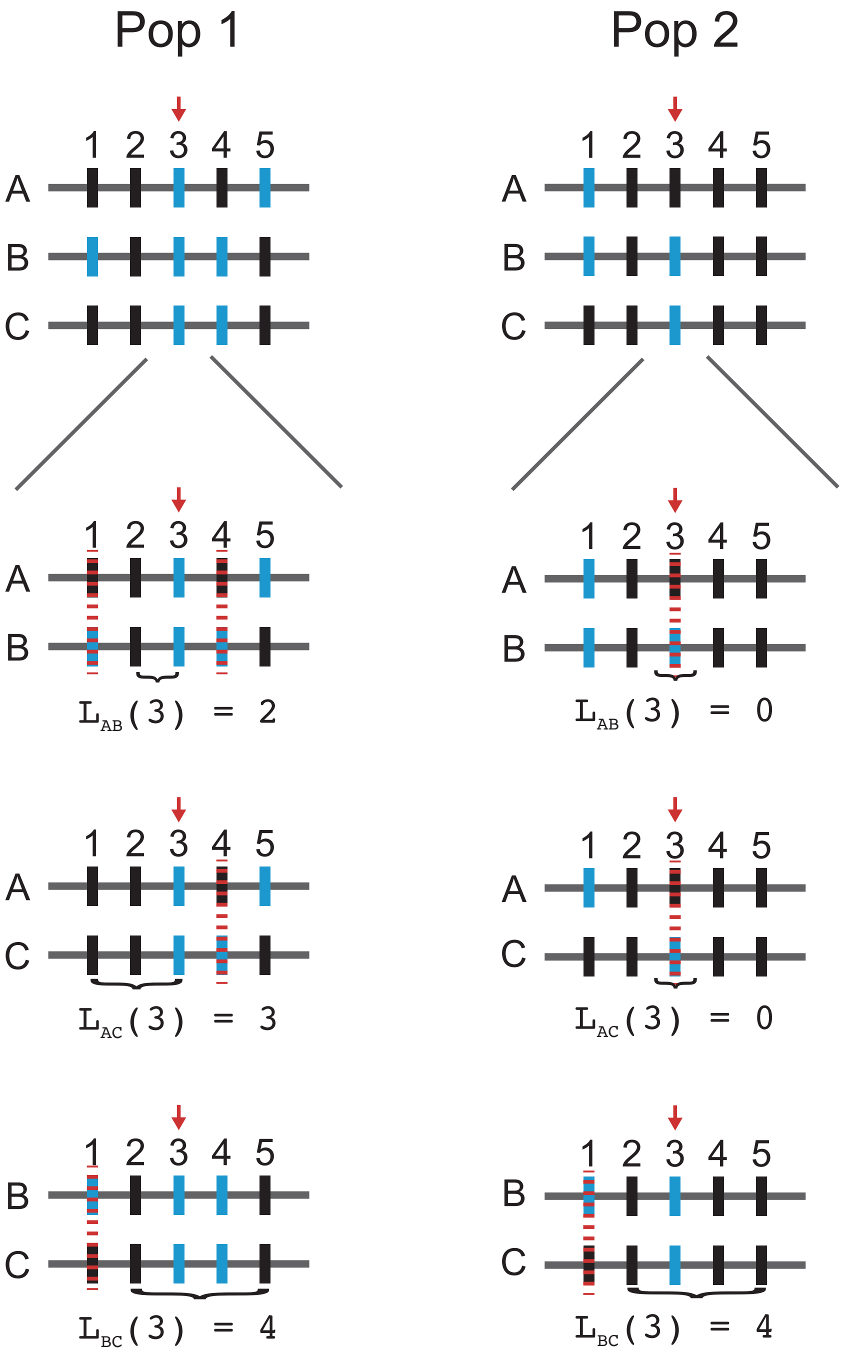


Figure 1. A representation of the demographic history for our high- and low-altitude populations as inferred by Liu *et al.* (2018).



$$SL_{P1}(3) = (2+3+4)/3 = 3$$

$$SL_{P2}(3) = (0+0+4)/3 = 4/3$$

$$XPnSL(3) = \log(SL_{P1}(3)/SL_{P2}(3)) = \log(9/4) = 0.3522$$

Figure 2. A toy example illustrating the computation of XP-nSL at a single site in two populations with three haplotypes (grey horizontal bars labeled A-C) and five sites (vertical bars labeled 1-5) where different alleles are colored blue or black. XP-nSL is calculated at site 3 (marked by red arrow). In each population, for each pair of haplotypes, the number of identical-by-state (IBS) sites are counted extending out from and including the test site (red arrow) until reaching a non-IBS site (marked by red dotted line). Within each population, the mean number of IBS sites is calculated across all pairs of haplotypes, and then the log-ratio of the mean from each population is computed to get XP-nSL at site 3.

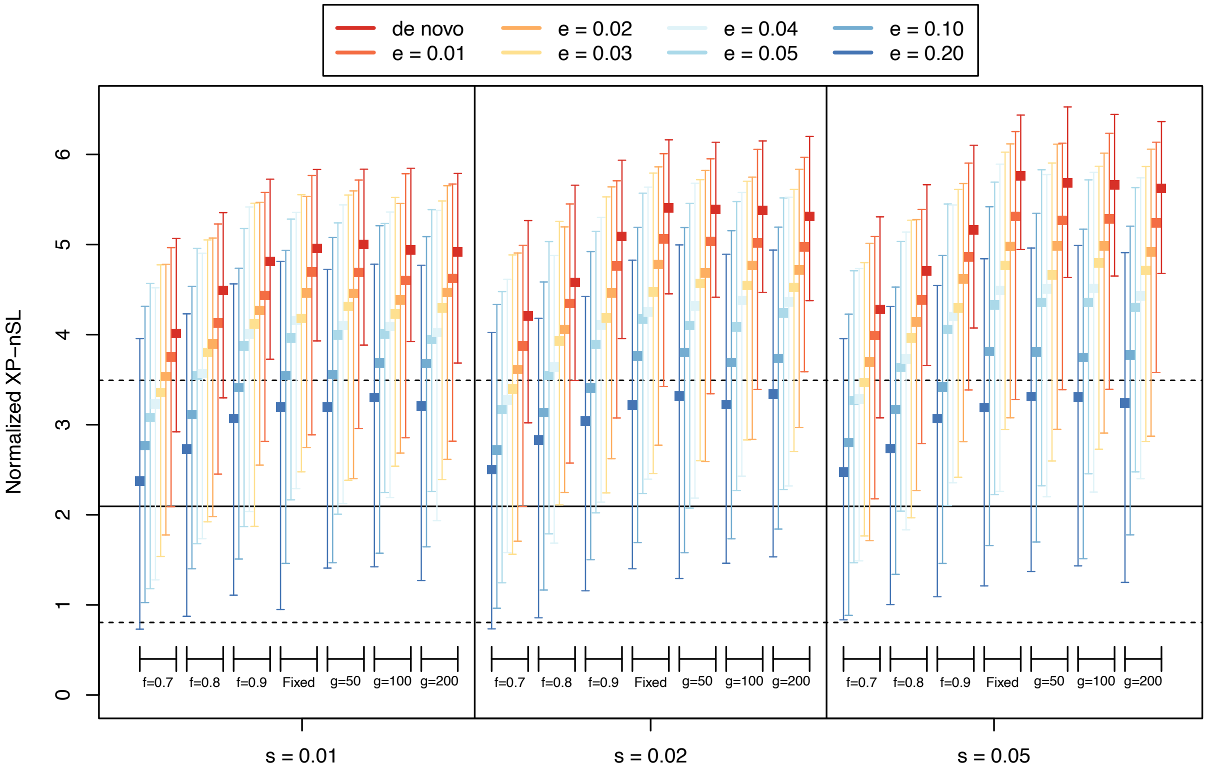


Figure 3. The distribution of maximum XP-nSL scores from simulations across various parameters, represented by medians and intervals containing 95% of the mass of the distribution. Neutral simulations are represented by the black solid horizonal line (median) and the black dashed horizonal line (95% interval). Non-neutral simulations represented by a colored box (median) and error bars (95% interval). The parameters are e (frequency at which selection begins, e > 0 indicates soft sweep), f (frequency of selected mutation at sampling), g (number of generations since fixation), and s (selection coefficient).

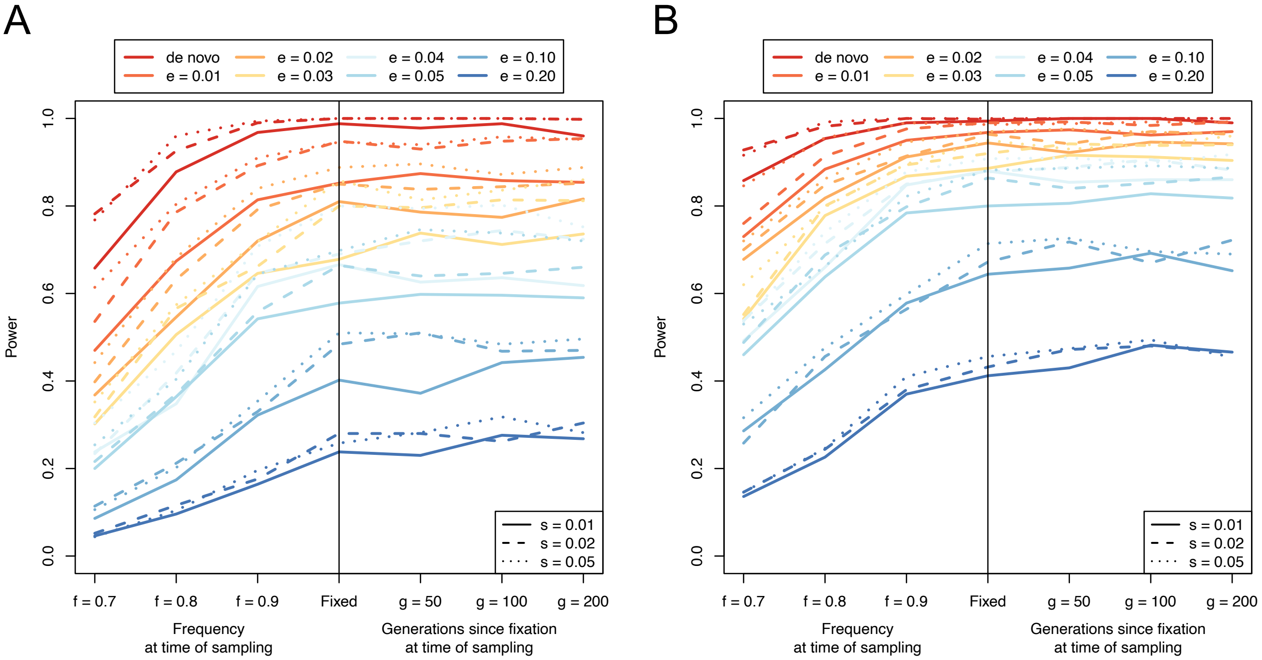
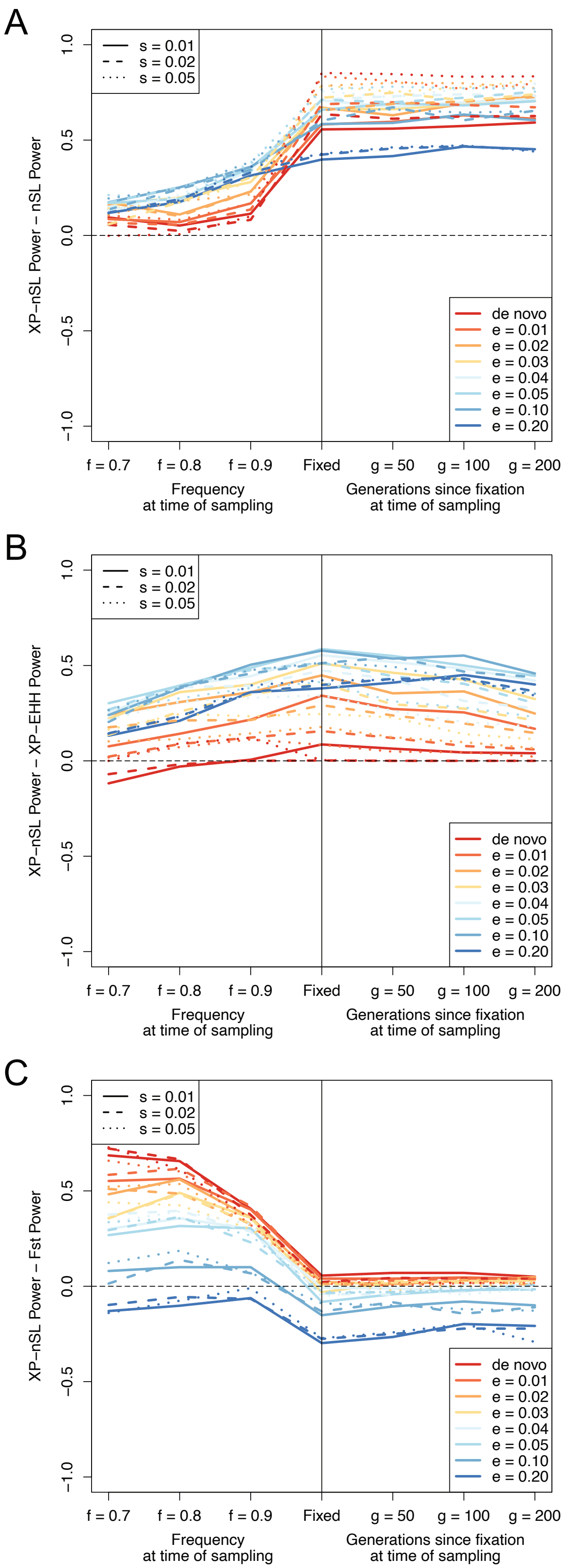


Figure 4. Power curves for (A) the max-score approach and (B) the window-based approach to identifying sweeps. The parameters are e (frequency at which selection begins, e > 0 indicates soft sweep), f (frequency of selected mutation at sampling), g (number of generations since fixation), and s (selection coefficient).


A
XP−nSL Power − nSL Power
B
XP−nSL Power − XP−EHH Power
C
XP−nSL Power − Fst Power
1.0
0.5
0.0
−0.5
−1.0
s = 0.01
s = 0.02
s = 0.05
de novo
e = 0.01
e = 0.02
e = 0.03
e = 0.04
e = 0.05
e = 0.10
e = 0.20
f = 0.7
f = 0.8
f = 0.9
Fixed
g = 50
g = 100
g = 200
Frequency
at time of sampling
Generations since fixation
at time of sampling

643

Figure 5. Difference in power between XP-nSL and (A) nSL, (B) XP-EHH, and (C) F_{ST}. Values above 0 indicate XP-nSL has more power, and values below 0 indicate XP-nSL has less power. The horizontal black dotted line marks 0. The parameters are e (frequency at which selection begins, e > 0 indicates soft sweep), f (frequency of selected mutation at sampling), g (number of generations since fixation), and s (selection coefficient).

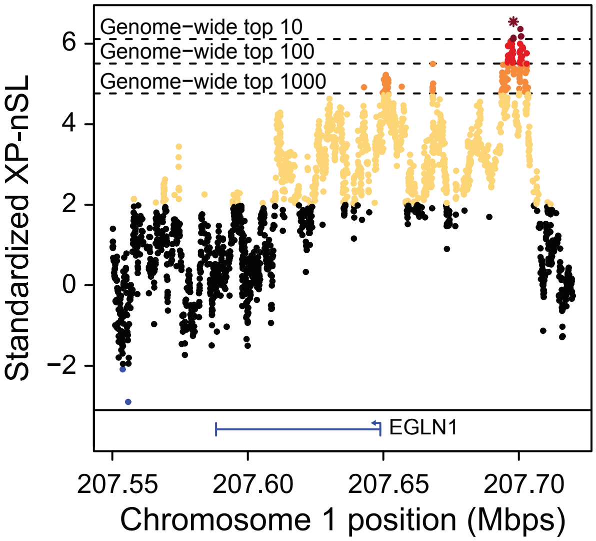


Figure 6. XP-nSL scores in the vicinity of the EGLN1 locus. This locus contains the genome wide top score (star) and six of the top ten genome wide scores (dark red).

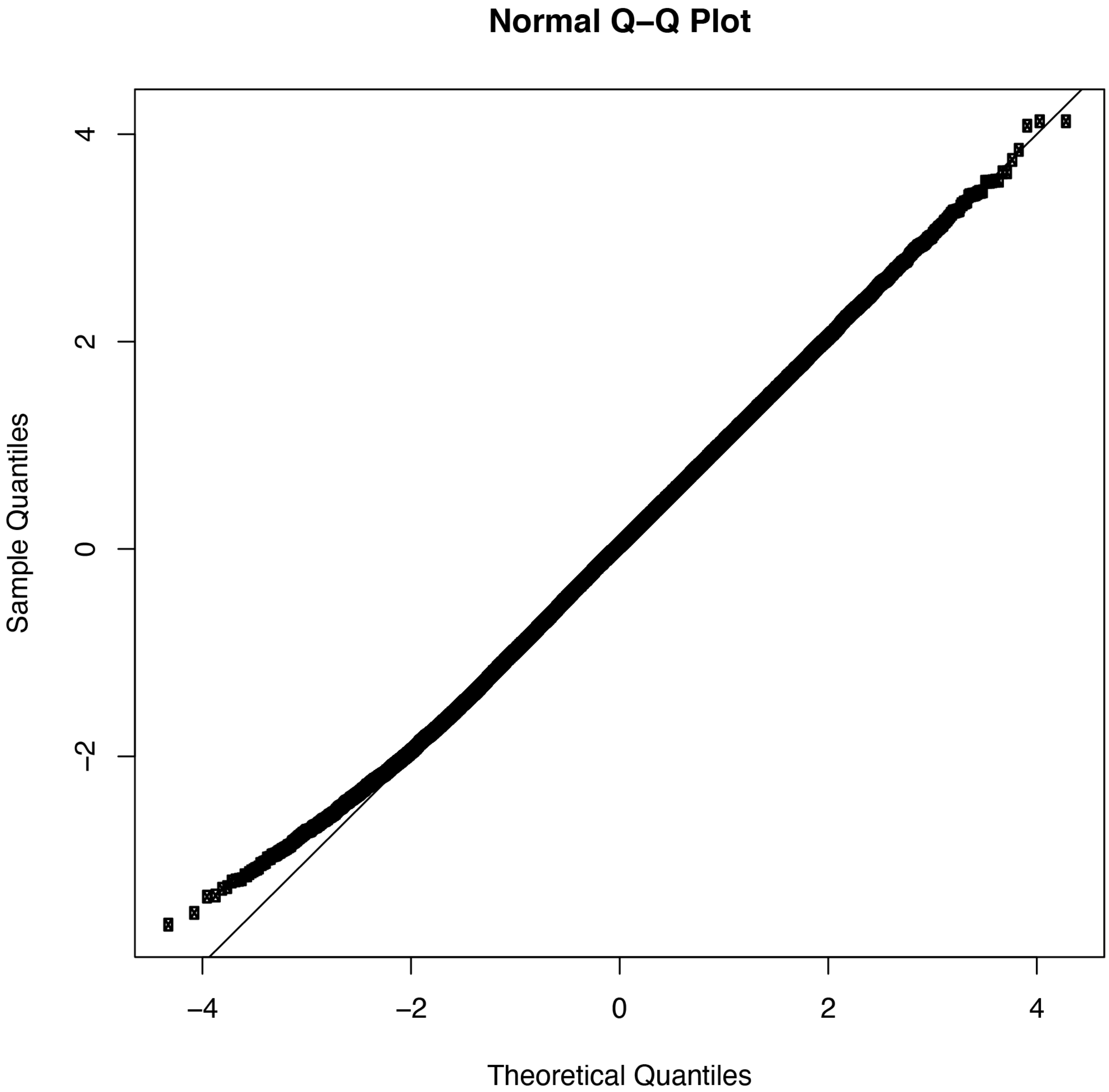


Figure S1. A normal quantile-quantile plot of neutral XP-nSL scores showing generally good adherence to a standard normal distribution. Due to autocorrelation along the genome, only every 1000th score is plotted.

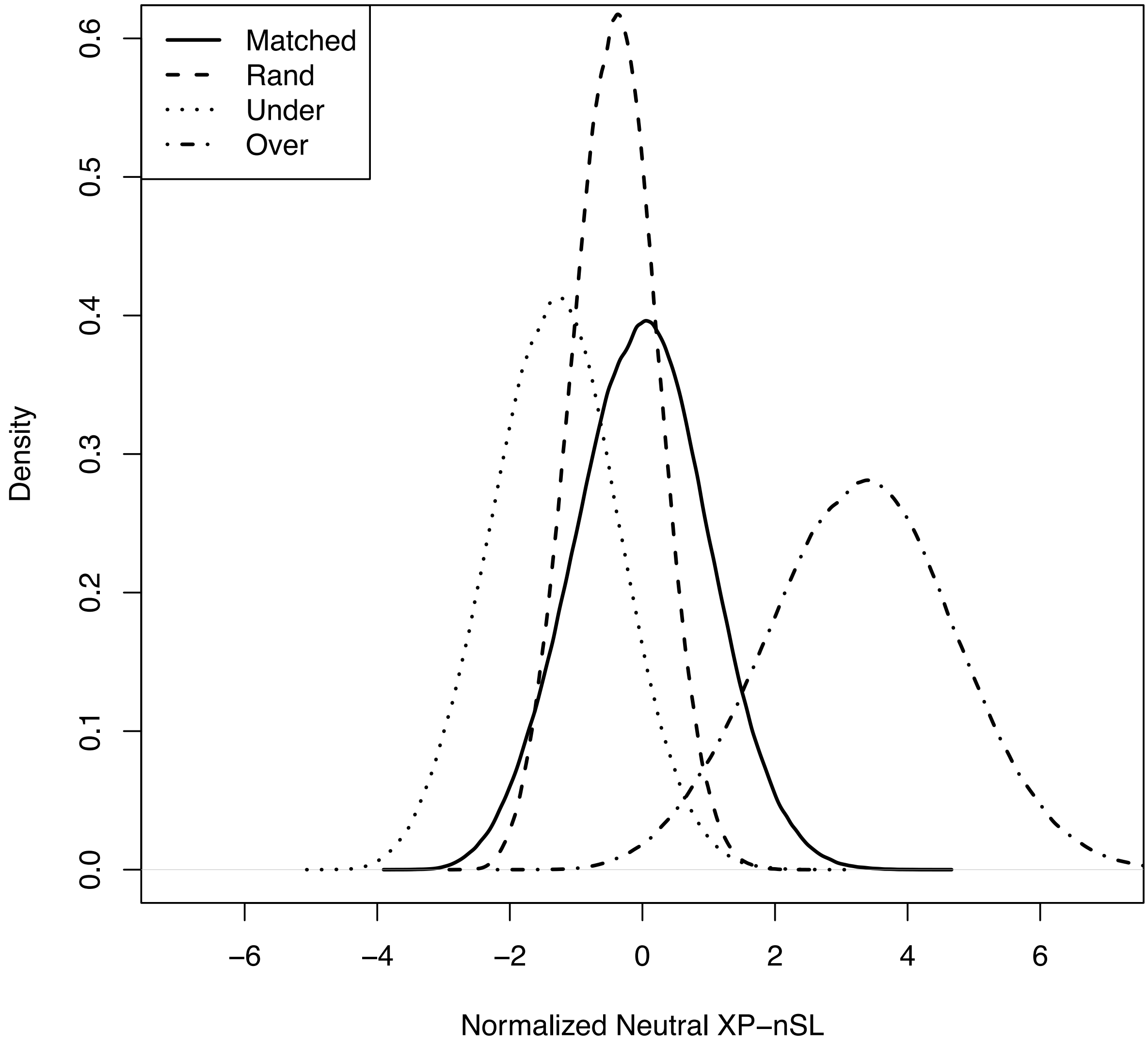


Figure S2. The distribution of neutral XP-nSL scores normalized with a matched demographic history (solid line), normalized with the “Rand” demographic history (dashed line), normalized with the “Under” demographic history (dotted line), and normalized with the “Over” demographic history (dash-dot line). Normalizing with the wrong demographic history can dramatically shift the distribution of neutral XP-nSL scores.

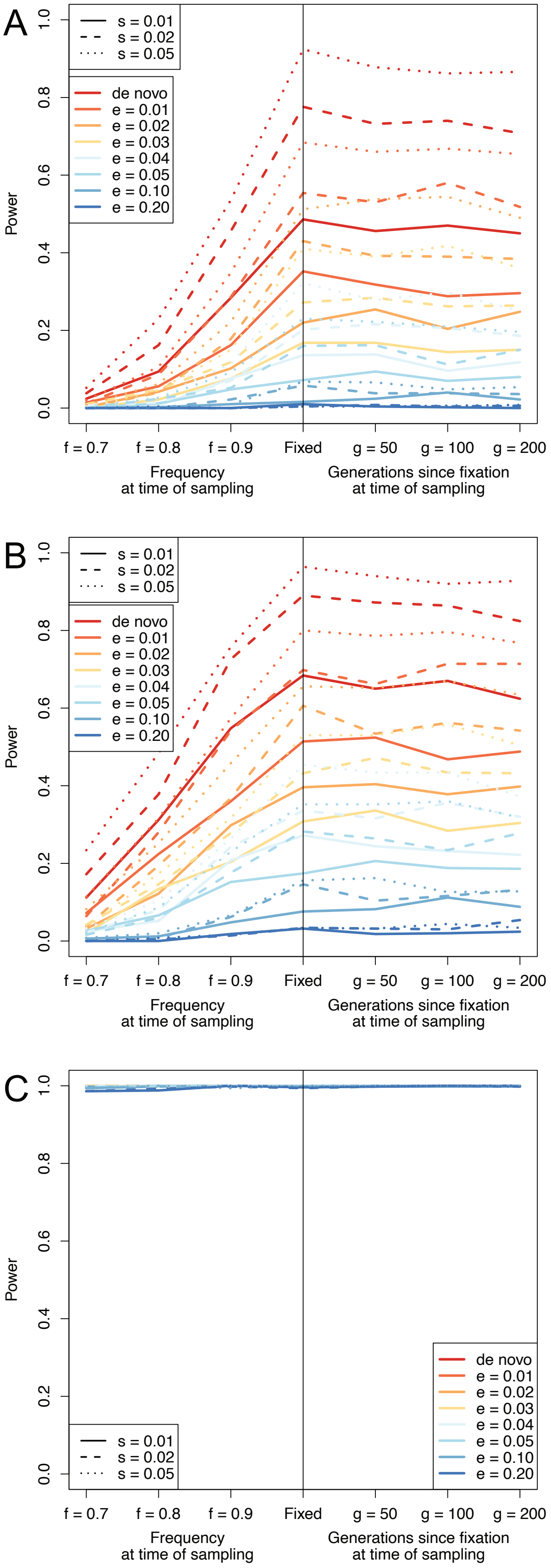

A
B
C
Power
0.0
0.2
0.4
0.6
0.8
1.0
s = 0.01
s = 0.02
s = 0.05
de novo
e = 0.01
e = 0.02
e = 0.03
e = 0.04
e = 0.05
e = 0.10
e = 0.20
f = 0.7
f = 0.8
f = 0.9
Fixed
g = 50
g = 100
g = 200
Frequency at time of sampling
Generations since fixation at time of sampling

Figure S3. XP-nSL power using mismatched demographic histories for normalization. (A) Using the “Rand” history. (B) Using the “Under” history. (C) Using the “Over” history.

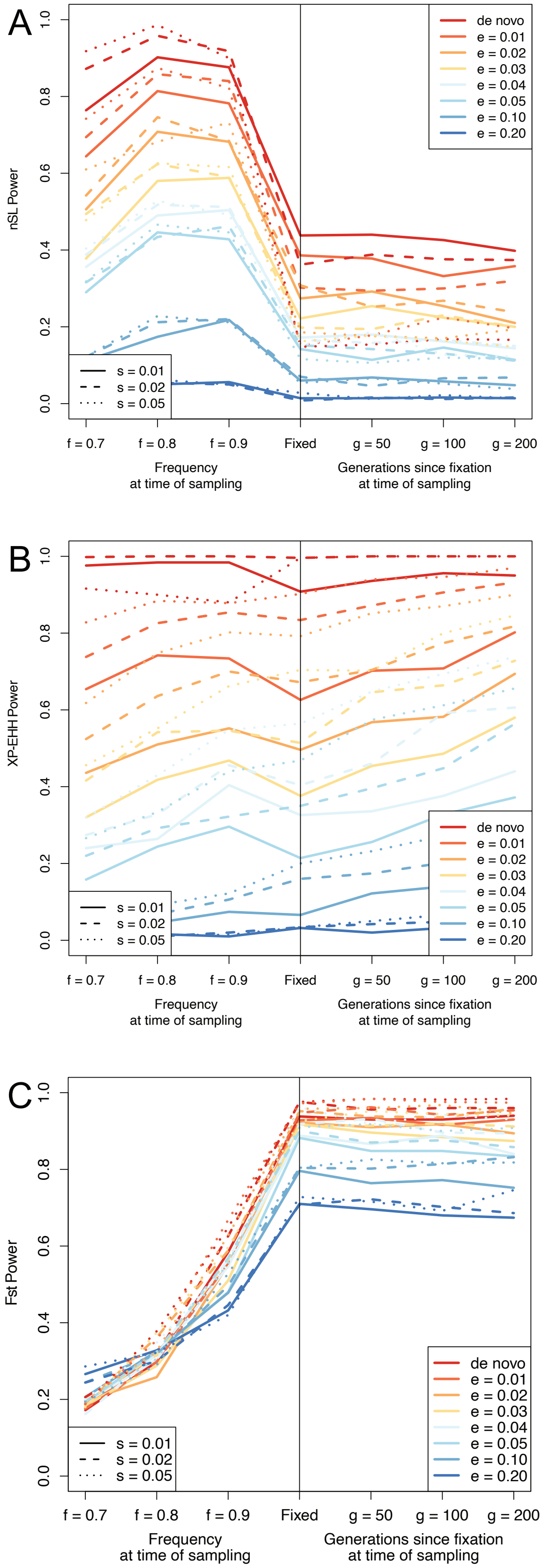

A
nSL Power
0.0
0.2
0.4
0.6
0.8
1.0
de novo
e = 0.01
e = 0.02
e = 0.03
e = 0.04
e = 0.05
e = 0.10
e = 0.20
s = 0.01
s = 0.02
s = 0.05
f = 0.7
f = 0.8
f = 0.9
Fixed
g = 50
g = 100
g = 200
Frequency
at time of sampling
Generations since fixation
at time of sampling
B
XP-EHH Power
0.0
0.2
0.4
0.6
0.8
1.0
de novo
e = 0.01
e = 0.02
e = 0.03
e = 0.04
e = 0.05
e = 0.10
e = 0.20
s = 0.01
s = 0.02
s = 0.05
f = 0.7
f = 0.8
f = 0.9
Fixed
g = 50
g = 100
g = 200
Frequency
at time of sampling
Generations since fixation
at time of sampling
C
Fst Power
0.0
0.2
0.4
0.6
0.8
1.0
de novo
e = 0.01
e = 0.02
e = 0.03
e = 0.04
e = 0.05
e = 0.10
e = 0.20
s = 0.01
s = 0.02
s = 0.05
f = 0.7
f = 0.8
f = 0.9
Fixed
g = 50
g = 100
g = 200
Frequency
at time of sampling
Generations since fixation
at time of sampling

Figure S4. Power curves for (A) nSL, (B) XP-EHH, and (C) F_{ST}. The parameters are e (frequency at which selection begins, e > 0 indicates soft sweep), f (frequency of selected mutation at sampling), g (number of generations since fixation), and s (selection coefficient).

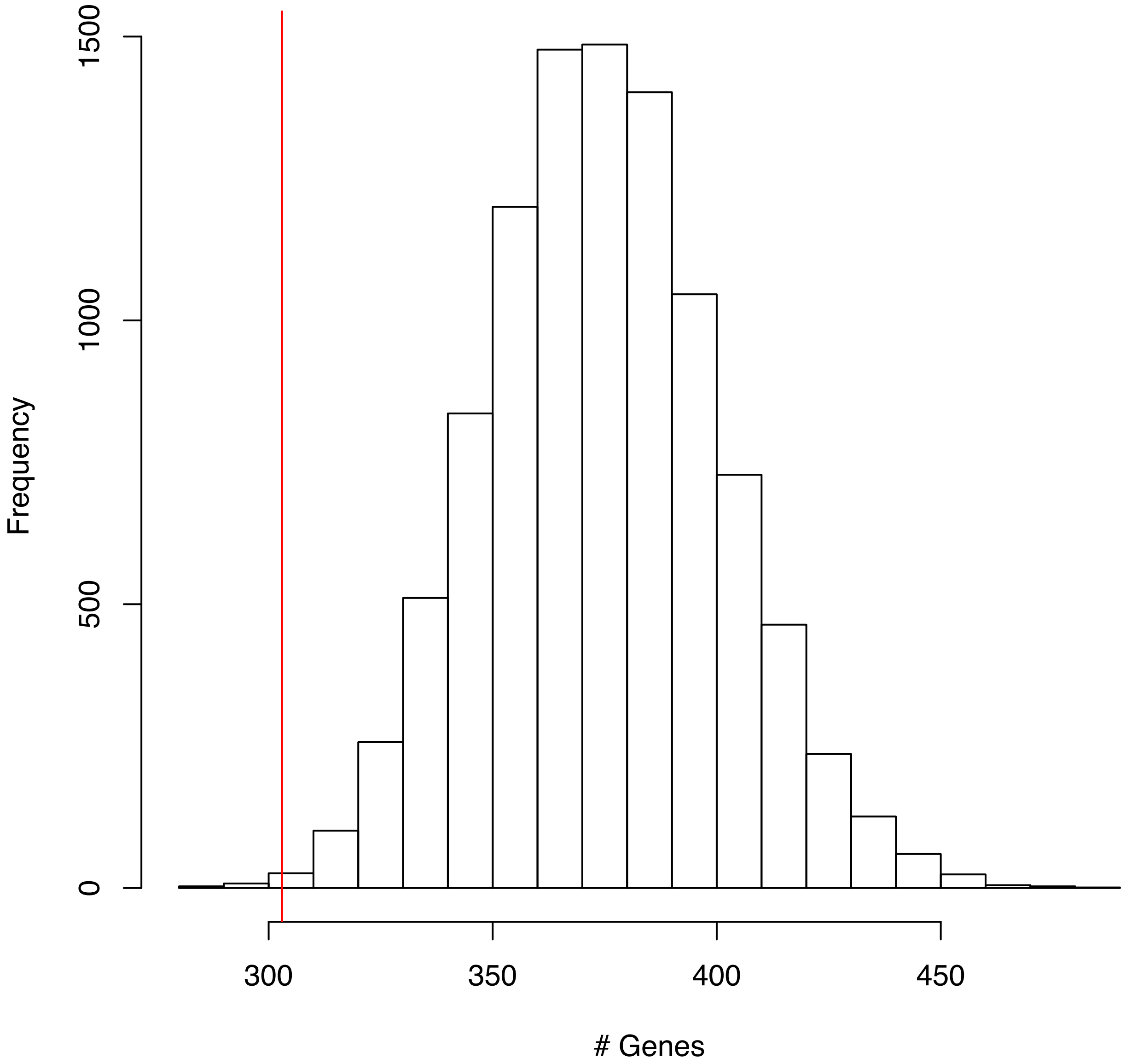


Figure S5. A permutation test (10,000 replicates) that shuffles 270 100kb regions around the macaque genome and counts the number of unique genes overlapping. The red vertical line marks the 303 genes found in the real data analysis. The probability of observing 303 or fewer genes is 1.4×10^{-3}, indicating the analysis is not randomly choosing gene regions.

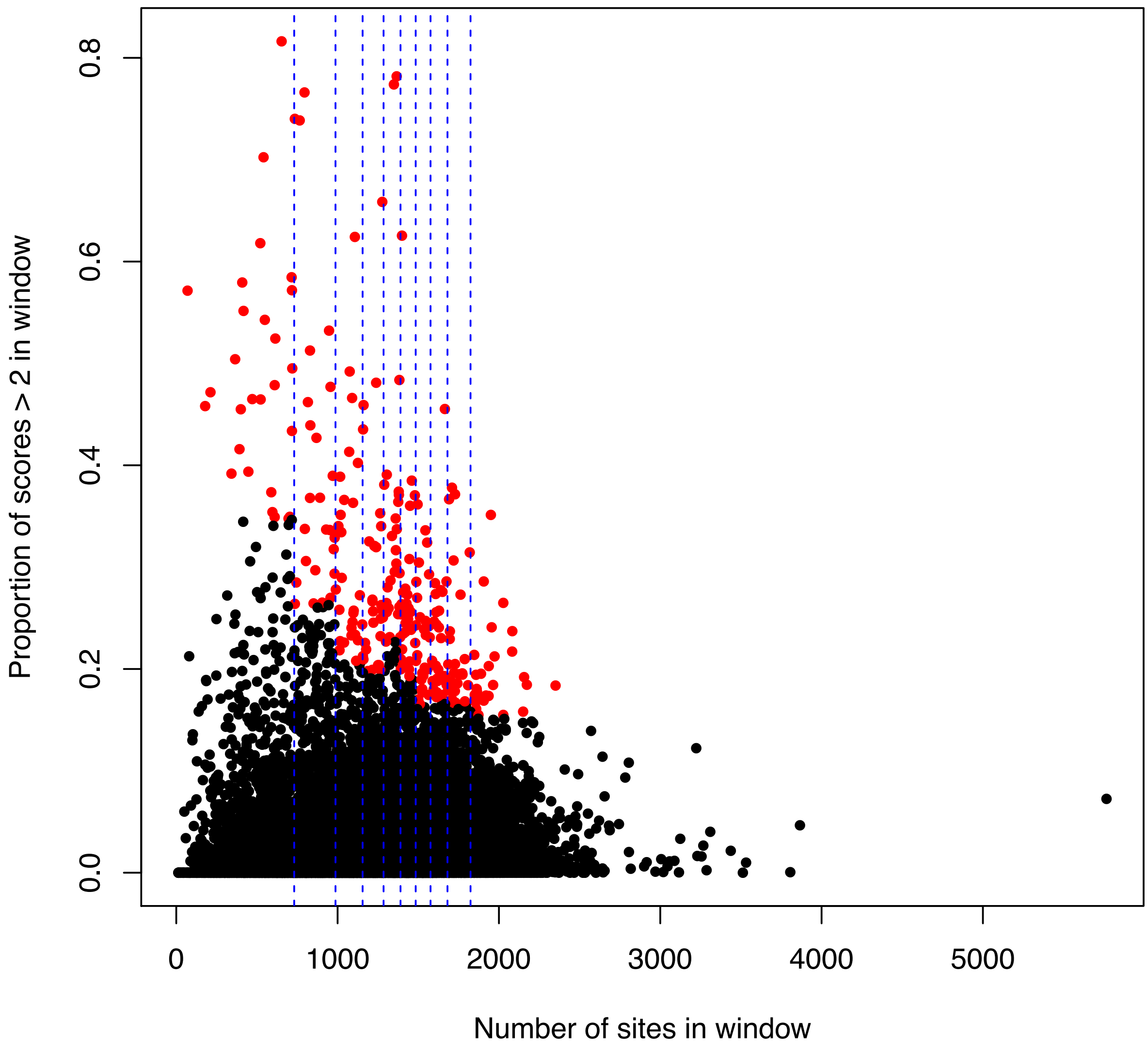


Figure S6. Proportion of scores > 2 versus number of sites in window. Blue vertical dashed lines indicate bin boundaries. Each circle is a window, red dots indicate a proportion of scores > 2 beyond the 1% threshold for that bin.

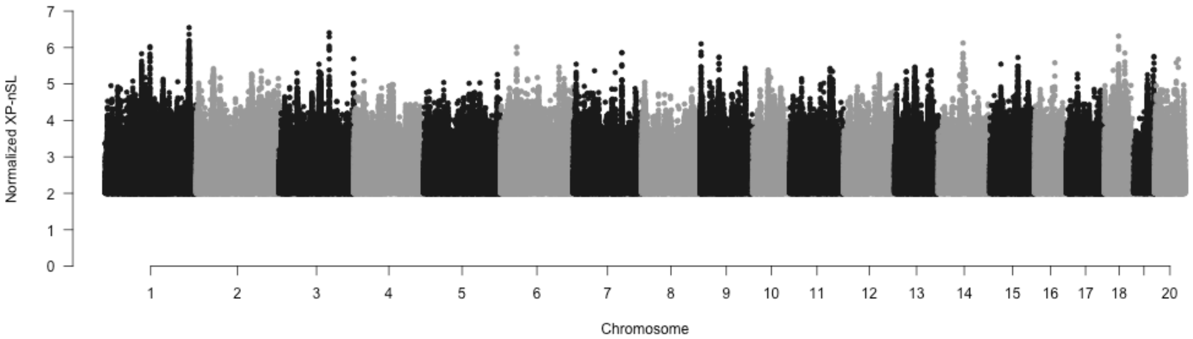


Figure S7. A Manhattan plot of normalized XP-nSL scores across the genome. Due to a very large number of points, only scores > 2 were plotted.

References

Aarts, M., Iihara, K., Wei, W.L., Xiong, Z.G., Arundine, M., Cerwinski, W. *et al.* (2003). A key role for TRPM7 channels in anoxic neuronal death. *Cell* 115**:** 863-877.

Ahmad, K.S., Hameed, M., Fatima, S., Ashraf, M., Ahmad, F., Naseer, M. *et al.* (2016). Morpho-anatomical and physiological adaptations to high altitude in some Aveneae grasses from Neelum Valley, Western Himalayan Kashmir. *Acta Physiologiae Plantarum* 38**:** 93.

Alston, C.L., Rocha, M.C., Lax, N.Z., Turnbull, D.M. & Taylor, R.W. (2017). The genetics and pathology of mitochondrial disease. *J Pathol* 241**:** 236-250.

BCM-HGSC. (2020) Baylor College of Medicine Human Genome Sequencing Center Rhemac8 Recombination Map. Available at: [ftp://ftp.hgsc.bcm.edu/ucscHub/rhesusSNVs/rheMac8/all.rate.bw].

Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J.M., Mei, R. *et al.* (2010). Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet* 6**:** e1001116.

Bigham, A.W. & Lee, F.S. (2014). Human high-altitude adaptation: forward genetics meets the HIF pathway. *Genes Dev* 28**:** 2189-2204.

Bigham, A.W., Mao, X., Mei, R., Brutsaert, T., Wilson, M.J., Julian, C.G. *et al.* (2009). Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. *Hum Genomics* 4**:** 79-90.

Cai, Q., Qian, X., Lang, Y., Luo, Y., Xu, J., Pan, S. *et al.* (2013). Genome sequence of ground tit Pseudopodoces humilis and its adaptation to high altitude. *Genome Biol* 14**:** R29.

Campbell, K.L., Storz, J.F., Signore, A.V., Moriyama, H., Catania, K.C., Payson, A.P. *et al.* (2010). Molecular basis of a novel adaptation to hypoxic-hypercapnia in a strictly fossorial mole. *BMC Evol Biol* 10**:** 214.

Cardenas, A., Villalba, A., de Juan Romero, C., Pico, E., Kyrousi, C., Tzika, A.C. *et al.* (2018). Evolution of Cortical Neurogenesis in Amniotes Controlled by Robo Signaling Levels. *Cell* 174**:** 590-606 e521.

Chappell, M.A. & Hammond, K.A. (2004). Maximal aerobic performance of deer mice in combined cold and exercise challenges. *J Comp Physiol B* 174**:** 41-48.

Cheviron, Z.A., Bachman, G.C., Connaty, A.D., McClelland, G.B. & Storz, J.F. (2012). Regulatory changes contribute to the adaptive enhancement of thermogenic capacity in high-altitude deer mice. *Proceedings of the National Academy of Sciences* 109**:** 8635-8640.

Cheviron, Z.A. & Brumfield, R.T. (2012). Genomic insights into adaptation to high-altitude environments. *Heredity* 108**:** 354-361.

Corcoran, A. & O'Connor, J.J. (2013). Hypoxia-inducible factor signalling mechanisms in the central nervous system. *Acta Physiol (Oxf)* 208**:** 298-310.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A. *et al.* (2011). The variant call format and VCFtools. *Bioinformatics* 27**:** 2156-2158.

Delaneau, O., Zagury, J.F., Robinson, M.R., Marchini, J.L. & Dermitzakis, E.T. (2019). Accurate, scalable and integrative haplotype estimation. *Nat Commun* 10**:** 5436.

Dewey, F.E., Gusarova, V., Dunbar, R.L., O'Dushlaine, C., Schurmann, C., Gottesman, O. *et al.* (2017). Genetic and Pharmacologic Inactivation of ANGPTL3 and Cardiovascular Disease. *N Engl J Med* 377**:** 211-221.

Fan, Z., Zhou, A., Osada, N., Yu, J., Jiang, J., Li, P. *et al.* (2018). Ancient hybridization and admixture in macaques (genus Macaca) inferred from whole genome sequences. *Mol Phylogenet Evol* 127**:** 376-386.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T. & Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol Biol Evol* 31**:** 1275-1291.

Fooden, J. (2000). *Systematic review of the rhesus macaque, Macaca mulatta (Zimmermann, 1780)*Field Museum of Natural History, Chicago, Ill. :.

Foti, A., Dorendorf, F. & Leimkuhler, S. (2017). A single nucleotide polymorphism causes enhanced radical oxygen species production by human aldehyde oxidase. *PLoS One* 12**:** e0182061.

Garud, N.R., Messer, P.W., Buzbas, E.O. & Petrov, D.A. (2015). Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLoS Genet* 11**:** e1005004.

Ge, R.-L., Cai, Q., Shen, Y.-Y., San, A., Ma, L., Zhang, Y. *et al.* (2013). Draft genome sequence of the Tibetan antelope. *Nature Communications* 4**:** 1858.

Gonzalo-Turpin, H. & Hazard, L. (2009). Local adaptation occurs along altitudinal gradient despite the existence of gene flow in the alpine plant species Festuca eskia. *Journal of Ecology* 97**:** 742-751.

Graham, A.M. & McCracken, K.G. (2019). Convergent evolution on the hypoxia-inducible factor (HIF) pathway genes EGLN1 and EPAS1 in high-altitude ducks. *Heredity (Edinb)* 122**:** 819-832.

Greinert, R., Volkmer, B., Henning, S., Breitbart, E.W., Greulich, K.O., Cardoso, M.C. *et al.* (2012). UVA-induced DNA double-strand breaks result from the repair of clustered oxidative DNA damages. *Nucleic Acids Res* 40**:** 10263-10273.

Guo, T., Yin, R.X., Huang, F., Yao, L.M., Lin, W.X. & Pan, S.L. (2016). Association between the DOCK7, PCSK9 and GALNT2 Gene Polymorphisms and Serum Lipid levels. *Sci Rep* 6**:** 19079.

Guo, X., Hu, Q., Hao, G., Wang, X., Zhang, D., Ma, T. *et al.* (2018). The genomes of two Eutrema species provide insight into plant adaptation to high altitudes. *DNA Research* 25**:** 307-315.

Hayes, J.P. & Chappell, M.A. (1986). Effects of Cold Acclimation on Maximum Oxygen Consumption during Cold Exposure and Treadmill Exercise in Deer Mice, Peromyscus maniculatus. *Physiological Zoology* 59**:** 473-481.

Hebbar, P., Nizam, R., Melhem, M., Alkayal, F., Elkum, N., John, S.E. *et al.* (2018). Genome-wide association study identifies novel recessive genetic variants for high TGs in an Arab population. *J Lipid Res* 59**:** 1951-1966.

Hendrickson, S.L. (2013). A genome wide study of genetic adaptation to high altitude in feral Andean Horses of the paramo. *BMC Evol Biol* 13**:** 273.

Hermisson, J. & Pennings, P.S. (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* 169**:** 2335-2352.

Hernandez, R.D., Hubisz, M.J., Wheeler, D.A., Smith, D.G., Ferguson, B., Rogers, J. *et al.* (2007). Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. *Science* 316**:** 240-243.

Huerta-Sanchez, E., Degiorgio, M., Pagani, L., Tarekegn, A., Ekong, R., Antao, T. *et al.* (2013). Genetic signatures reveal high-altitude adaptation in a set of ethiopian populations. *Mol Biol Evol* 30**:** 1877-1888.

Huerta-Sanchez, E., Jin, X., Asan, Bianba, Z., Peter, B.M., Vinckenbosch, N. *et al.* (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512**:** 194-197.

Ismail, I.H., Gagne, J.P., Genois, M.M., Strickfaden, H., McDonald, D., Xu, Z. *et al.* (2015). The RNF138 E3 ligase displaces Ku to promote DNA end resection and regulate DNA repair pathway choice. *Nat Cell Biol* 17**:** 1446-1457.

Jandl, K., Thekkekara Puthenparampil, H., Marsh, L.M., Hoffmann, J., Wilhelm, J., Veith, C. *et al.* (2019). Long non-coding RNAs influence the transcriptome in pulmonary arterial hypertension: the role of PAXIP1-AS1. *J Pathol* 247**:** 357-370.

Janocha, A.J., Comhair, S.A.A., Basnyat, B., Neupane, M., Gebremedhin, A., Khan, A. *et al.* (2017). Antioxidant defense and oxidative damage vary widely among high-altitude residents. *Am J Hum Biol* 29.

Jeong, C., Alkorta-Aranburu, G., Basnyat, B., Neupane, M., Witonsky, D.B., Pritchard, J.K. *et al.* (2014). Admixture facilitates genetic adaptations to high altitude in Tibet. *Nat Commun* 5**:** 3281.

Karle, C., Gehrig, T., Wodopia, R., Hoschele, S., Kreye, V.A., Katus, H.A. *et al.* (2004). Hypoxia-induced inhibition of whole cell membrane currents and ion transport of A549 cells. *Am J Physiol Lung Cell Mol Physiol* 286**:** L1154-1160.

Kern, A.D. & Schrider, D.R. (2016). Discoal: flexible coalescent simulations with selection. *Bioinformatics* 32**:** 3839-3841.

Kim, E. & Donohue, K. (2013). Local adaptation and plasticity of Erysimum capitatum to altitude: its implications for responses to climate change. *Journal of Ecology* 101**:** 796-805.

Kim, Y. & Nielsen, R. (2004). Linkage Disequilibrium as a Signature of Selective Sweeps. *Genetics* 167**:** 1513-1524.

Krueger, S.K. & Williams, D.E. (2005). Mammalian flavin-containing monooxygenases: structure/function, genetic polymorphisms and role in drug metabolism. *Pharmacol Ther* 106**:** 357-387.

Kumar, R., Corbett, M.A., Smith, N.J., Jolly, L.A., Tan, C., Keating, D.J. *et al.* (2015). Homozygous mutation of STXBP5L explains an autosomal recessive infantile-onset neurodegenerative disorder. *Hum Mol Genet* 24**:** 2000-2010.

Li, M., Tian, S., Jin, L., Zhou, G., Li, Y., Zhang, Y. *et al.* (2013). Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat Genet* 45**:** 1431-1438.

Li, Y., Wu, D.D., Boyko, A.R., Wang, G.D., Wu, S.F., Irwin, D.M. *et al.* (2014). Population variation revealed high-altitude adaptation of Tibetan mastiffs. *Mol Biol Evol* 31**:** 1200-1205.

Liu, J.-Q., Duan, Y.-W., Hao, G., Ge, X.-J. & Sun, H. (2014). Evolutionary history and underlying adaptation of alpine plants on the Qinghai–Tibet Plateau. *Journal of Systematics and Evolution* 52**:** 241-249.

Liu, Z., Tan, X., Orozco-terWengel, P., Zhou, X., Zhang, L., Tian, S. *et al.* (2018). Population genomics of wild Chinese rhesus macaques reveals a dynamic demographic history and local adaptation, with implications for biomedical research. *Gigascience* 7.

Madrid, J.E., Mandalaywala, T.M., Coyne, S.P., Ahloy-Dallaire, J., Garner, J.P., Barr, C.S. *et al.* (2018). Adaptive developmental plasticity in rhesus macaques: the serotonin transporter gene interacts with maternal care to affect juvenile social behaviour. *Proc Biol Sci* 285.

Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P.D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* 47**:** D419-D426.

Munne-Bosch, S., Cotado, A., Morales, M., Fleta-Soriano, E., Villellas, J. & Garcia, M.B. (2016). Adaptation of the Long-Lived Monocarpic Perennial Saxifraga longifolia to High Altitude. *Plant Physiol* 172**:** 765-775.

O'Reilly, P.F., Birney, E. & Balding, D.J. (2008). Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Res* 18**:** 1304-1313.

Osada, N. & Akashi, H. (2012). Mitochondrial-nuclear interactions and accelerated compensatory evolution: evidence from the primate cytochrome C oxidase complex. *Mol Biol Evol* 29**:** 337-346.

Peng, Y., Yang, Z., Zhang, H., Cui, C., Qi, X., Luo, X. *et al.* (2011). Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Mol Biol Evol* 28**:** 1075-1081.

Pennings, P.S. & Hermisson, J. (2006). Soft sweeps II--molecular population genetics of adaptation from recurrent mutation or migration. *Mol Biol Evol* 23**:** 1076-1084.

Przeworski, M. (2002). The Signature of Positive Selection at Randomly Chosen Loci. *Genetics* 160**:** 1179-1189.

Qiu, Q., Zhang, G., Ma, T., Qian, W., Wang, J., Ye, Z. *et al.* (2012). The yak genome and adaptation to life at high altitude. *Nat Genet* 44**:** 946-949.

Qu, Y., Zhao, H., Han, N., Zhou, G., Song, G., Gao, B. *et al.* (2013). Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau. *Nat Commun* 4**:** 2071.

Ramasamy, S.K., Kusumbe, A.P., Wang, L. & Adams, R.H. (2014). Endothelial Notch activity promotes angiogenesis and osteogenesis in bone. *Nature* 507**:** 376-380.

Richard, A.F., Goldstein, S.J. & Dewar, R.E. (1989). Weed macaques: The evolutionary implications of macaque feeding ecology. *International Journal of Primatology* 10**:** 569.

Rodriguez, J., Pilkington, R., Garcia Munoz, A., Nguyen, L.K., Rauch, N., Kennedy, S. *et al.* (2016). Substrate-Trapped Interactors of PHD3 and FIH Cluster in Distinct Signaling Pathways. *Cell Rep* 14**:** 2745-2760.

Rosenmann, M. & Morrison, P. (1974). Maximum oxygen consumption and heat loss facilitation in small homeotherms by He-O2. *Am J Physiol* 226**:** 490-495.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F. *et al.* (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419**:** 832-837.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C. *et al.* (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449**:** 913-918.

Schweizer, R.M., Velotta, J.P., Ivy, C.M., Jones, M.R., Muir, S.M., Bradburd, G.S. *et al.* (2019). Physiological and genomic evidence that selection on the transcription factor Epas1 has altered cardiovascular function in high-altitude deer mice. *PLoS Genet* 15**:** e1008420.

Shang, P., Li, W., Liu, G., Zhang, J., Li, M., Wu, L. *et al.* (2019). Identification of lncRNAs and Genes Responsible for Fatness and Fatty Acid Composition Traits between the Tibetan and Yorkshire Pigs. *Int J Genomics* 2019**:** 5070975.

Simonson, T.S. (2015). Altitude Adaptation: A Glimpse Through Various Lenses. *High Alt Med Biol* 16**:** 125-137.

Stewart, C.B. & Disotell, T.R. (1998). Primate evolution - in and out of Africa. *Curr Biol* 8**:** R582-588.

Storz, J.F. & Cheviron, Z.A. (2021). Physiological Genomics of Adaptation to High-Altitude Hypoxia. *Annu Rev Anim Biosci* 9**:** 149-171.

Storz, J.F., Cheviron, Z.A., McClelland, G.B. & Scott, G.R. (2019). Evolution of physiological performance capacities and environmental adaptation: insights from high-elevation deer mice (Peromyscus maniculatus). *J Mammal* 100**:** 910-922.

Storz, J.F., Sabatino, S.J., Hoffmann, F.G., Gering, E.J., Moriyama, H., Ferrand, N. *et al.* (2007). The molecular basis of high-altitude adaptation in deer mice. *PLoS Genet* 3**:** e45.

Szpiech, Z.A. & Hernandez, R.D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol* 31**:** 2824-2827.

Tanaka, T., Inazawa, J. & Nakamura, Y. (1996). Molecular cloning and mapping of a human cDNA for cytosolic malate dehydrogenase (MDH1). *Genomics* 32**:** 128-130.

To, K.K. & Huang, L.E. (2005). Suppression of hypoxia-inducible factor 1alpha (HIF-1alpha) transcriptional activity by the HIF prolyl hydroxylase EGLN1. *J Biol Chem* 280**:** 38102-38107.

Velotta, J.P., Robertson, C.E., Schweizer, R.M., McClelland, G.B. & Cheviron, Z.A. (2020). Adaptive shifts in gene regulation underlie a developmental delay in thermogenesis in high-altitude deer mice. *Molecular Biology and Evolution*.

Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. (2006). A map of recent positive selection in the human genome. *Plos Biol* 4**:** e72.

Wang, G.D., Fan, R.X., Zhai, W., Liu, F., Wang, L., Zhong, L. *et al.* (2014). Genetic convergence in the adaptation of dogs and humans to the high-altitude environment of the tibetan plateau. *Genome Biol Evol* 6**:** 2122-2128.

Wang, M.-S., Li, Y., Peng, M.-S., Zhong, L., Wang, Z.-J., Li, Q.-Y. *et al.* (2015). Genomic Analyses Reveal Potential Independent Adaptation to High Altitude in Tibetan Chickens. *Molecular Biology and Evolution* 32**:** 1880-1889.

Wang, M.-S., Wang, S., Li, Y., Jhala, Y., Thakur, M., Otecko, N.O. *et al.* (2020). Ancient hybridization with an unknown population facilitated high altitude adaptation of canids. *Molecular Biology and Evolution*.

Wang, X., Takenaka, K. & Takeda, S. (2010). PTIP promotes DNA double-strand break repair through homologous recombination. *Genes Cells* 15**:** 243-254.

Weir, B.S. & Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution* 38**:** 1358-1370.

Witt, K.E. & Huerta-Sanchez, E. (2019). Convergent evolution in human and domesticate adaptation to high-altitude environments. *Philos Trans R Soc Lond B Biol Sci* 374**:** 20180235.

Xing, J., Wang, M., Hong, J., Gao, Y., Liu, Y., Gu, H. *et al.* (2019). TRPM7 channel inhibition exacerbates pulmonary arterial hypertension through MEK/ERK pathway. *Aging (Albany NY)* 11**:** 4050-4065.

Xu, S., Li, S., Yang, Y., Tan, J., Lou, H., Jin, W. *et al.* (2010). A Genome-Wide Search for Signals of High-Altitude Adaptation in Tibetans. *Molecular Biology and Evolution* 28**:** 1003-1011.

Xu, Z., Jin, X., Cai, W., Zhou, M., Shao, P., Yang, Z. *et al.* (2018). Proteomics Analysis Reveals Abnormal Electron Transport and Excessive Oxidative Stress Cause Mitochondrial Dysfunction in Placental Tissues of Early-Onset Preeclampsia. *Proteomics Clin Appl* 12**:** e1700165.

Yang, X., Wang, Y., Zhang, Y., Lee, W.H. & Zhang, Y. (2016). Rich diversity and potency of skin antioxidant peptides revealed a novel molecular basis for high-altitude adaptation of amphibians. *Sci Rep* 6**:** 19866.

Ye, B., Hou, N., Xiao, L., Xu, Y., Xu, H. & Li, F. (2016). Dynamic monitoring of oxidative DNA double-strand break and repair in cardiomyocytes. *Cardiovasc Pathol* 25**:** 93-100.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X., Pool, J.E. *et al.* (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329**:** 75-78.

Zhang, Y., Yuan, F., Wu, X., Rechkoblit, O., Taylor, J.S., Geacintov, N.E. *et al.* (2000). Error-prone lesion bypass by human DNA polymerase eta. *Nucleic Acids Res* 28**:** 4717-4724.