Supplemental Material

A flexible method for estimating the fraction of fitness influencing mutations from large sequencing data sets

Sunjin Moon & Joshua M. Akey

Table of Contents

Supplemental Notes

1. Simulation Models

1.1 Neutral mutations and population expansion

Demographic scenario: After splitting into two populations, one population experiences rapid growth to 10,000,000, and other population remains at the same population size (10,000) before the split (Gazave et al. 2013). The mutation rate per site is varied under the demographic model.

Command line of sfs_code

./sfs_code 2 1 -N 1000 -t %f -n 5000 -a N -W 0 -Td 0 P 0 0.0758 -Td 0.02 P 0 13.2 -Td 0.2 P 0 0.0769 -Td 0.207 P 0 13 -TS 0.215 0 1 -Tg 0.216 P 1 345.37501 -TE 0.236

Where: %f token depicts substitution for θ.

θ = (0.0016, 0.0008, 0.0004)

1.2 Selection and realistic demographic model

Demographic scenario involving European demographic model (Tennessen et al. 2012; Gazave et al. 2013): After splitting into two populations, one population experiences a slow recovery from severe population bottleneck, followed by exponential growth, N_e = 512,010. The other population's N_e remains constant at 9,210.

Command line of sfs_code:

./sfs_code 2 1000 -N 740 -t 0.008 -r 0.0008 -n %d -W %s -TN 0 1447 -TN 0.2622 186 -TN 0.3378 104 -TS 0.3379 0 1 -Tg 0.3379 45.47 -Tg 0.3861 P 0 0 -Tg 0.3861 P 1 283.16 -TE 0.40

Where, %d and %s tokens depict substitutions for sample size and selection coefficients.

a) Sample size

%d = (5, 50, 500, 5000)

b) Selection coefficients

i) Neutral mutation %s : '0'

ii) Purifying selection (%s):

-|ϒ|: "L 0 2 0 0.0 1.0 0.206 0.0004931"

-|ϒ|*0.1: "L 0 2 0 0.0 1.0 0.206 0.004931"

-|ϒ|*0.025: "L 0 2 0 0.0 1.0 0.206 0.0197240"

-|ϒ|*0.01: "L 0 2 0 0.0 1.0 0.206 0.04931"

-|ϒ|*0.005: "L 0 2 0 0.0 1.0 0.206 0.0986200"

-|ϒ|*0.0025: "L 0 2 0 0.0 1.0 0.206 0.1972400"

-|ϒ|*0.001: "L 0 2 0 0.0 1.0 0.206 0.4931"

-|ϒ|*0.0001: "L 0 2 0 0.0 1.0 0.206 4.931"

iii) Positive selection (%s):

|ϒ|: "L 0 2 1.0 0.206 0.0004931 0.206 0.0004931"

|ϒ|*0.1: "L 0 2 1.0 0.206 0.004931 0.206 0.004931"

|ϒ|*0.025: "L 0 2 1.0 0.206 0.0197240 0.206 0.0197240"

|ϒ|*0.01: "L 0 2 1.0 0.206 0.0493100 0.206 0.0493100"

|ϒ|*0.005: "L 0 2 1.0 0.206 0.0986200 0.206 0.0986200"

|ϒ|*0.0025: "L 0 2 1.0 0.206 0.1972400 0.206 0.1972400"

|ϒ|*0.001: "L 0 2 1.0 0.206 0.4931 0.206 0.4931"

|ϒ|*0.0001: "L 0 2 1.0 0.206 4.931 0.206 4.931"

iv) Mixture of negative selection and positive selection (%s)

-|ϒ|*0.1 and |ϒ|*0.1 : "L 0 1 41.77652 %n %p"

where, %n = (0.10, 0.20, …, 1.00) is the proportion of sites under purifying selection, %p = (0.0, 0.02,…, 0.10) is the proportion of sites under positive selection, and %n + %p = 1. The same gamma distribution is used for both negative and positive selection.

Note: ϒ is the 'baseline' model of distribution of selection coefficients as inferred by Boyko et al (Boyko et al. 2008), where $\alpha = 0.206$ and $\beta = 1/2740$. Under each combination of demographic parameters and selection coefficients, 100 sets of sequences were generated for compiling SFS and computing the mean and variance of f estimates.

1.3 Simulation of exon sequences

A set of SFS was constructed from 1,000 simulation replicates. Each replicate consisted of 300bp of coding sequence (e.g., synonymous and nonsynonymous sites) and 100bp of non-coding sequence. Positions of non-coding sequences were 0, 1kb, 10kb, 100kb, 500kbp from coding regions (**Supplemental Figure 1**). For selection models, only nonsynonymous sites were under selection with a selection coefficient drawn form the given Gamma distribution. Non-coding sequences were used as the reference SFS, unless otherwise noted.

2 Data analysis

2.1 Building site-frequency spectrum

We initially separated coding SNVs into genomic contexts based on CDS annotations of Homo sapiens (GRch37.74). We chose the longest CDS if alternative isoforms existed for a gene. For estimating codon usage of synonymous sites, we used codon frequencies based on tRNAscan_SE Analysis of Homo sapiens (hg19 – NCBI Build 37.1) that can be found in http://gtrnadb.ucsc.edu/Hsapi19/. PhyloP scores (Cooper et al. 2005) of 99 vertebrate genomes were used for measuring evolutionary conservation that can be found at the UCSC genome annotation database (human build GRCh37/hg19). Regulatory potentials of SNVs were identified based on the datasets of high-resolution maps of DnaseI Hypersensitive Sites (DHS) identified in 81 human cell types by using DNaseI FDR-1% calls for genome-wide per-nucleotide DNaseI-seq (Stergachis et al. 2013). The datasets were downloaded from the UW ENCODE Project (http://www.uwencode.org/proj/Science_Stergachis_et_al/). Levels of local recombination rates were defined by using a quantile binning procedure in which the same number of data points fall within each group (i.e., top 5% quantile) as recombination hot-spots and bottom 5% as recombination cold-spots, from the distribution of recombination rates across the human genome. We used the combined HapMap recombination map that is available at the UCSC genome browser. Then, we separated intronic SNVs located within 50bp of each exon into the same context class. Finally, we defined the unfolded SFS as a vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_{n-1})$, where η_i denotes the number of sites that consisted of i derived alleles and n-i ancestral alleles for a given context.

2.2 Smoothing the SFS

When the sample size is large and sequence length is small, the SFS can have many zero intermediate frequency categories, which may result in inaccurate estimation of θ-estimators. We applied the one-dimensional smoothing spline fit to the SFS by using '*interpolante.UnivariateSpline'* method in the Python library "scipy" library to obtain robust estimation of θ-estimator. The python library for estimating f can be found in https://github.com/moon-s/fraction-under-selection.

2.3 Python code for estimating *f*

```
#!/usr/bin/python
#

import sys
import numpy as np
import scipy.interpolate as interp
import operator as op


# input SFS
# sfs is 1-D arrary of site frequency(1,2,..., n-1)
#

def ncr(n, r):
    r = min(r, n-r)
    if r == 0: return 1
    numer = reduce(op.mul, xrange(n, n-r, -1))
    denom = reduce(op.mul, xrange(1, r+1))
```

```
    return numer//denom


def theta_w(sfs):
          N = len( sfs ) + 1
          a1 = np.array(xrange(1, N ))
          return sum( sfs)/sum(1.0/a1) # sum 1 ~ N-1 class


def theta_pi(sfs):
          N = len( sfs )  + 1
          s = sum( i*( N - i )*sfs[ i -1 ] for i in range( 1, N) )
          return s/float( ncr(N, 2 ) )



def diff_f( test, ref ):
          f = 0.0
          # optional: smoothing
          n = len( test )
          x = np.array(range(1, len(  test)  +1))
          best_s = int( n * np.var(test))
          i_test = interp.UnivariateSpline (x, test, s = best_s)
          best_s = int( n * np.var(ref))
          i_ref = interp.UnivariateSpline (x, ref , s= best_s)
          sfs_test = i_test( x )
          sfs_ref  = i_ref( x )
          # scaling SFS_test by ratio of theta_pi
          a1 = theta_pi( sfs_test)/theta_pi( sfs_ref )
          selectiontype = "negative "
          f = ( 1 - theta_w(sfs_ref*a1)/theta_w(sfs_test) )
          if  f < 0.:
                    # scale by theta_w
                    selectiontype = "positive"
                    s = np.asarray( [i*(n+1 - i) for i in range(1, n + 1 )] )/float(ncr(n+1, 2 ))
                    a2 = theta_w( sfs_test)/theta_w(sfs_ref )
                    f = -( 1 - theta_pi( ref*a2) /theta_pi( test) )
          return selectiontype , f


#
# load SFS from files into array
#
if len( sys.argv ) == 3 :
          file_testsfs = open( sys.argv[1] )
          file_refsfs = open( sys.argv[2])
          testsfs, refsfs = (), ()
          for l in file_testsfs:
                    testsfs = np.array([ int(x) for x in l.split()] )
          for l in file_refsfs:
                    refsfs = np.array([ int(x) for x in l.split() ])
          print  "Fraction of sites under {} selection: {}".format( *diff_f( testsfs, refsfs )[0:2] )
```

3 Evaluation of θ-estimators in calculating f

3.1 Choice of θ-estimators

Out of various θ-estimators, we used θ_W and θ_π as estimators of θ, where $\theta = 4N\mu$ (N is effective population size, μ is mutation rate), as summary statistics of the scaled population mutation rate. Under the standard neutral model, θ_π and θ_W are equal to one another, but are differentially influenced by rare and common alleles. Due to these characteristics, we used θ_π when testing for positive selection and θ_W when testing for purifying selection. The theoretical variance of θ_W is:

$E(\theta_W) = a_1 M$,

$V(S) = a_1 M + a_2 M^2$, where

$a_1=\sum_1^{n-1} 1/i$, $a_2 = \sum_1^{n-1} 1/i^2$, and M = 4Nμ.

Theoretical variance of θ_π is:

$E(\theta_\pi) = M$,

$V(\theta_\pi) = b_1M + b_2M^2$, where

$b_1= \frac{n+1}{c(n-1)}$, $b_2 = \frac{2(n^2+n+3)}{9n(n-1)}$.

Analytical results show their variance decreases as the sample size increases (**Supplemental Figure S2**). This property contributes to stable estimation of *f*.

3.2 θ-estimators for scaling factor

Under the neutral mutation model, assuming a common demographic history in two sets of sites, θ-estimator would be the only parameter that characterizes the quantitative differences the two SFSs. Consistent with this expectation, there was no difference between the shape of test and reference SFS under various demographic perturbations (**Figure 1c,d**), except mutation rate. Based on this property, we calculated the ratio of θ in test sites to that in reference sites and used this to rescale the reference SFS. For a given SFS, $\boldsymbol{\eta} = \eta_i$ (i = 1, 2,…, n-1), where η_i is the number of sites at *i* derived allele frequency, common θ-estimators include:

$\theta_W = \frac{1}{a_1}\sum_{i=1}^{n-1}\eta_i$, $a_1= \sum_{i=1}^{n-1} 1/i$

$\theta\pi= \binom{2}{n}^{-1}\sum_{i=1}^{n-1} i(n-1)\eta_i$

$\theta_H= \binom{2}{n}^{-1}\sum_{i=1}^{n-1} i^2\eta_i$

$\theta_S = \eta_1$.

θ_W (Watterson 1975) and θ_π (Tajima 1983) have equal weight across site frequency categories and high weight on rare allele, respectively. θ_H (Fay and Wu 2000) has high weight on high frequency categories. θ_S (Fu and Li 1993) is the number of singletons. Both θ_H and θ_S require ancestral state information, whereas θ_π and θ_W do not. Additionally, we tested other possible θ-estimators, including the partial sum of $\boldsymbol{\eta}$, e.g., $S_1= \sum_{j=1}^{n/2}\eta_i$ and $\theta S_2 =\sum_{j=n/2}^{0.8*n}\eta_i$, for calculating the scaling factor, where S_1 is the sum of sites from singleton to intermediate frequency classes and S_2 is the sum of sites at low frequency and high frequency categories. Regardless of the choice of θ-estimators, we obtain $E(f)= \sum_{i=1}^{n-1}(\boldsymbol{\eta}_{i,test} - \alpha\boldsymbol{\eta}_{i,ref}) = 0$, where $\boldsymbol{\eta}$ is neutral site frequency spectrum, and $\alpha = \theta_{test}/\theta_{ref}$ (**Supplemental Figure S2**).

4 Factors influencing estimation accuracy

4.1 Robustness to hitchhiking, background selection, and recombination rate heterogeneity

Positive or purifying selection would act on closely linked sites by hitchhiking or background selection,

respectively. We tested these effects on estimates of *f* by simulating a coding region, consisting of neutral loci (synonymous sites) and loci subject to the selection (nonsynonymous sites), and a reference non-coding region, where the distance between coding region and non-coding region was varied (**Supplemental Figure 1**). We test *f* for synonymous variants with noncoding variants as a reference group. For close distance between them, both hitchhiking and background selection have no effect on the estimate of *f* (**Supplemental Figure S5**). For hitchhiking model, the estimate of *f* increases as the distance between them increases, reflecting hitchhiking effect produces an excess of rare variants at the linked neutral sites (Braverman et al. 1995). Meanwhile background selection, which reduces θ at the liked neutral sites (Stephan 2010), had no effect on the estimate of *f*. Essentially, recombination rate is a key evolutionary parameter that determines the degree of those effects at linked sites. We further tested potential effects of recombination rate heterogeneity on estimates of *f* by binning variants according to their local recombination rates (**Supplemental Figure 3**). We find no effect of recombination rate on estimates of *f*.

4.2 Effect of genomic contexts

We considered two genomic contexts, including CpG and GC-biased gene conversion. CpG sites are hypermutable for the spontaneous deamination of methylated cytosine at CpG to thymine (Ying and Huttley 2011), producing an excess of rare alleles. GC-biased gene conversion increases the rate of fixation of C/G alleles over A/T alleles (Capra et al. 2013), producing an excess of common alleles. Qualitatively, the neutral mutation rate determines the total amount of variants for a given sequence length (**Supplemental Figure 6**), but does not influence the SFS. In contrast, genomic context can influence the shape of the SFS because of fixation rate heterogeneity that might be confounded with selection pressures. We tested how estimates of *f* are affected by the mismatch of genomic contexts between test and reference sites (**Supplemental Figure 7**). As expected, the estimate of *f* was reduced for a combination of gBGC sites in the test model and NCB sites in the reference model, suggesting gBGC sites inflate common allele frequency by high fixation rate. Thus, matching genomic context between test and reference sites is critical to disentangle selection from neutral forces that influence the SFS.

4.3 Influence of selection on reference sites

Approximately 2.3% of intronic sites are under constraint (Pollard et al. 2010). Indeed, we found that estimates of *f* increased by removing highly conserved sites from reference sites (**Supplemental Figure 9**). Thus, estimates of *f* are conservative (i.e., biased downward) in the presence of deleterious sites in the reference set. To mitigate the potential influence of selection on reference sites in empirical analyses, we filtered out intronic sites that were in the top and bottom 5^{th} percentile of the distribution of PhyloP scores of intronic variants.

5 Empirical measure of selection strength

5.1 Selection strength and sample size

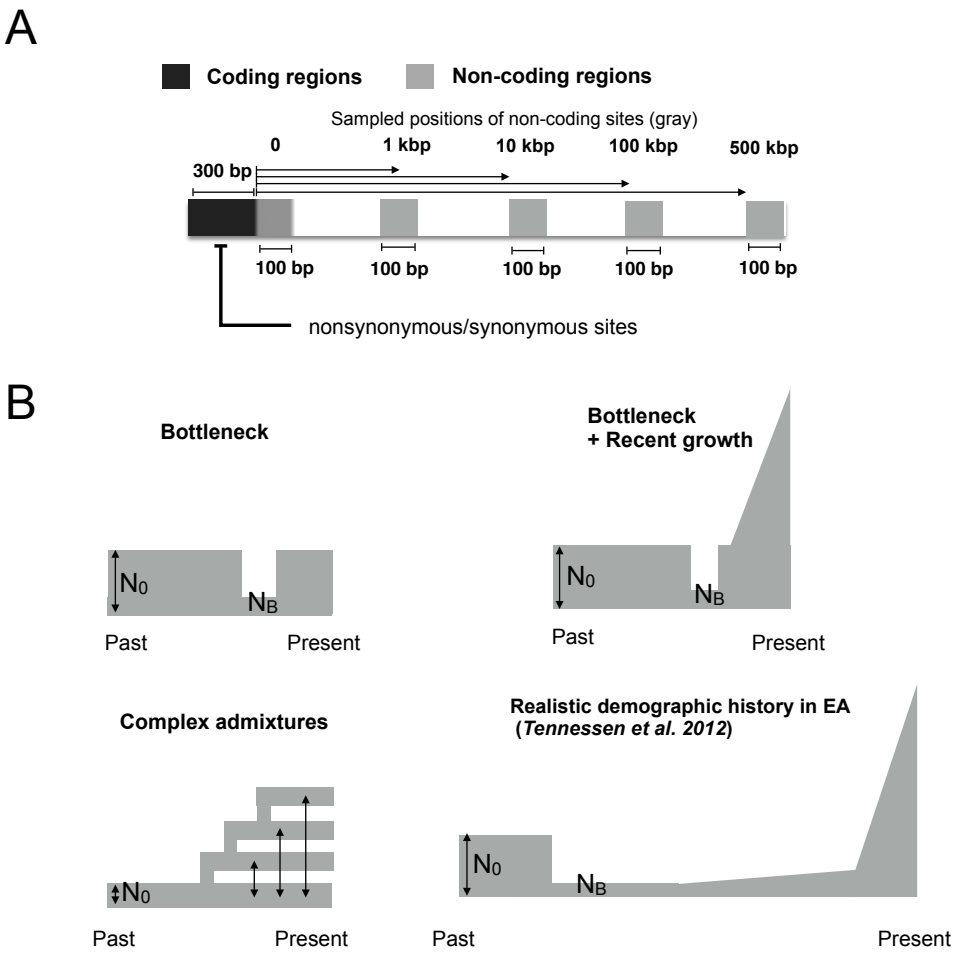
The abundance of rare alleles is a signature of explosive population growth (Keinan and Clark 2012). We investigated the effect of sample size on estimates of *f* under a model of rapid population expansion with constant selection coefficient. The estimate of *f* was increased as the sample size increased (**Supplemental Figure S4**), suggesting large samples are necessary to capture low frequency deleterious alleles under very strong selection. With large enough sample sizes, the estimate of *f* would reflect the observable fraction of deleterious alleles and be correlated with the strength of selection. For estimates of *f* larger than 0.6, where mutations under very strong purifying selection are quickly lost from the population, caution is needed in relating estimates of *f* to selection strength without prior knowledge of demographic history.

5.2 Mixture of different types of selections

For a mixture model of selection, we investigated a situation in which a fraction of mutations is deleterious and a fraction is advantageous, and, the reminders are neutral. We found that the estimates of *f* were sum of *f*(positive selection) + *f*(negative selection) + *f*(positive selection)×*f*(negative selection) (**Supplemental Figure S12**). In humans, the overall proportion of positively selected mutations in coding regions was expected to be less than 2%(Boyko et al. 2008). The genome-wide estimate of *f* will be reduced to the fraction of sites under positive selection. Practically, without a prior knowledge of which selection regime acted on test sites, a conservative interpretation of *f* would be the total fraction of observable non-neutral variants that some fraction of sites under one selection regime could be offset by the fraction of sites under the opposite selection regime.

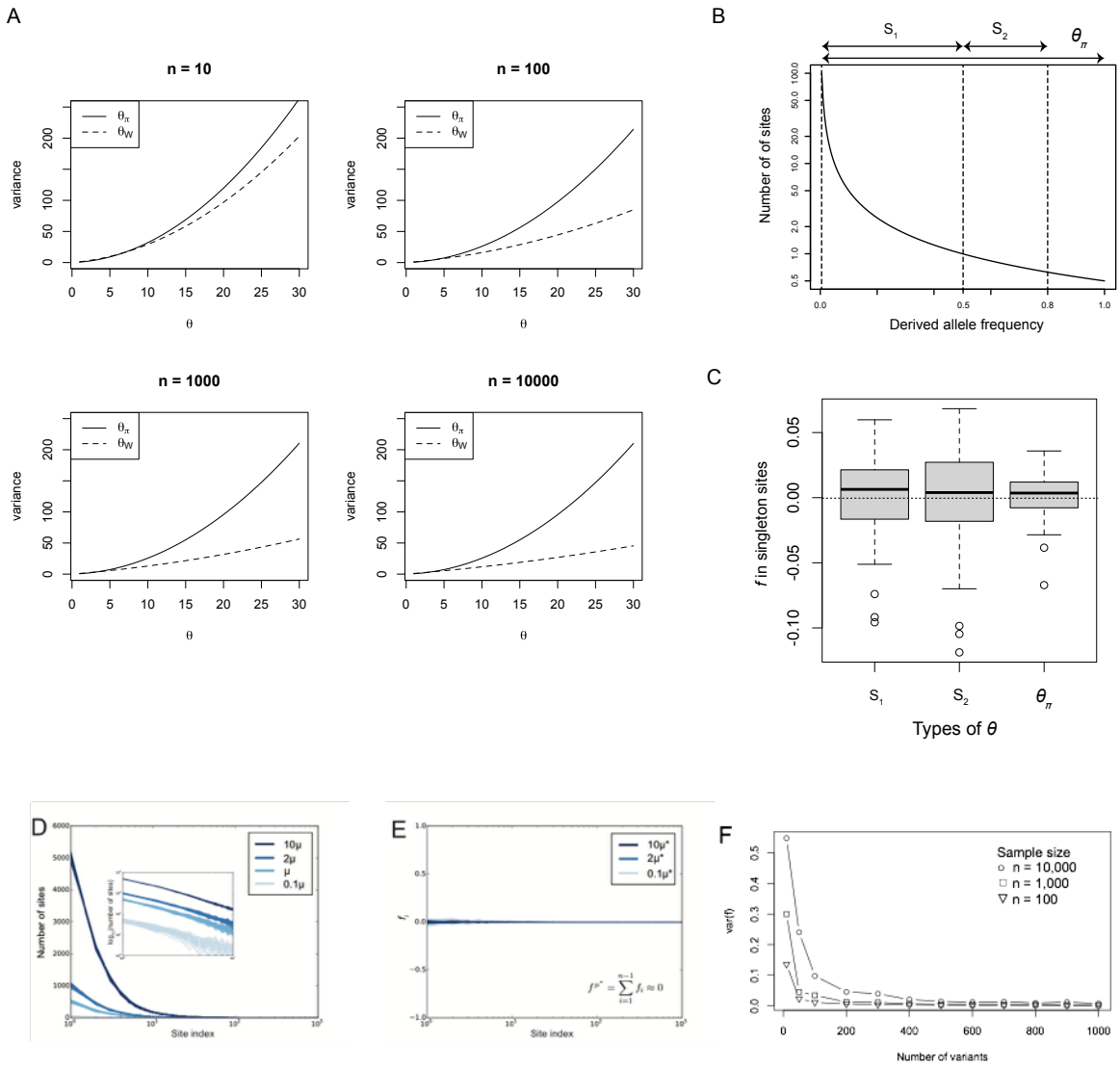
Supplemental Figures and Legends

Supplemental Figure S1



Supplemental Figure S1. Simulation model under complex demographic scenarios. (A) Simulated sequence consisted of a coding region (nonsynonymous and synonymous sites) and noncoding regions. For distributing the computational burden among the grid resources, 1,000 replicates under each condition were used for one SFS. A standard error was estimated from a set of hundred SFS for a given set of parameters. (B) Diagrams of demographic models used in the forward-time simulation. N_0 and N_B depict effective population size of ancestral population and effective population size during bottleneck period, respectively. All simulations were carried out using the program SFS-Code.

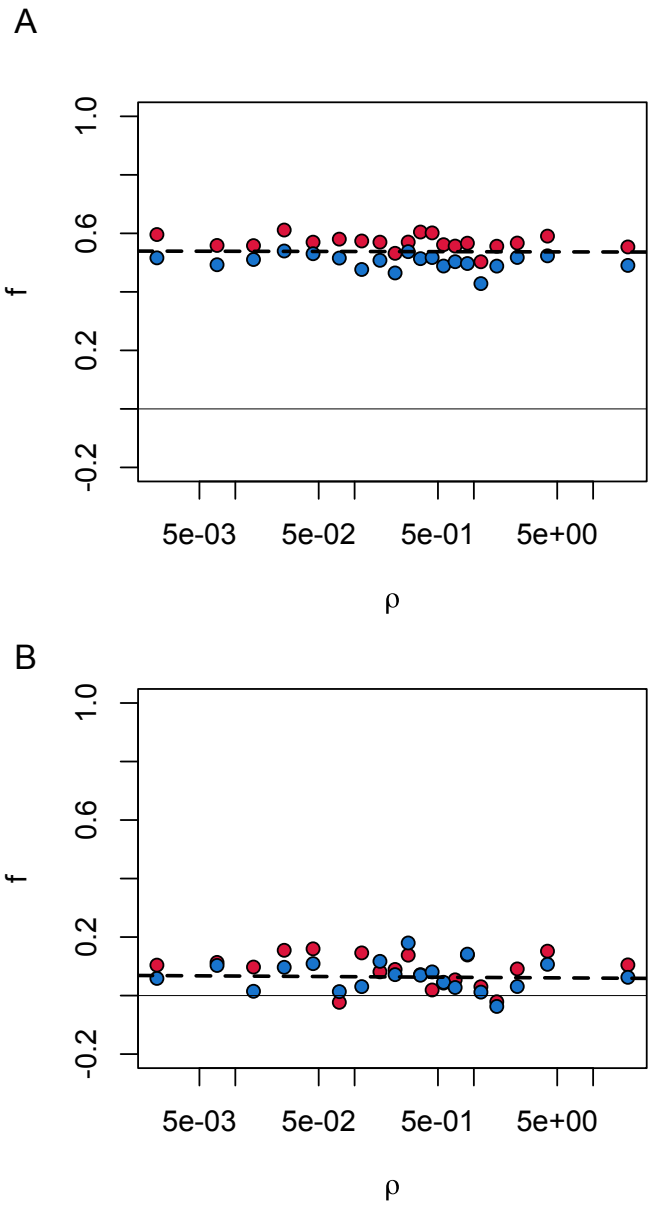
Supplemental Figure S2



Supplemental Figure S2. Variance of θ-estimators. (A) Theoretical variance of θ-estimators as a function of sample size (n). (B) Characteristics of summary statistics of θ in terms of site frequency spectrum. S_1 and S_2 are partial sums of the number of variable sites, where S_1 is the sum for rare to intermediate frequency sites, $\sum_{i=1}^{n/2} \eta_i$, and S_2 is the truncated sum for intermediate frequency sites, $\sum_{i=n/2}^{0.8*n} \eta_i$. (C) For two models of neutral variants under a realistic demographic model (n = 10,000), where one has a higher mutation rate than the other, differences in the rare frequency variation between them became approximately zero after scaling down the SFS with high mutation rate to have the same θ with the other SFS. (D) SFS of neutral mutations simulated under a realistic model of human demographic history (Tennessen et al. 2012; Gazave et al. 2013) with varying mutations rates. Each line corresponds to a different simulation replicate. (E) The scaled difference in SFS between the reference

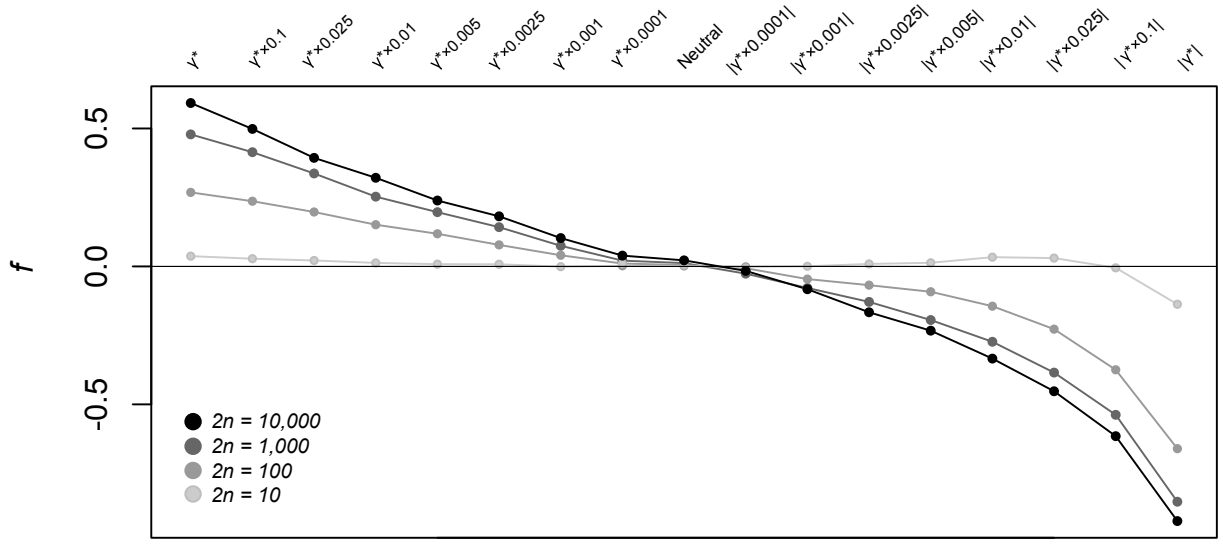
SFS (μ) and test SFS with a higher (2μ,10μ) or lower (μ/10) neutral mutation rate for each frequency class. Note, although the test and reference SFS have different mutation rates, the summed value of f across all frequency classes is zero. (G) Variance of the estimate of f for a realistic model of selection, obtained from the Gamma distribution of selection coefficients for nonsynonymous sites estimated in humans (Boyko et al. 2008), as a function of the number of variants. Since the variance of estimators of θ depended on the variance of each site frequency category (Ramirez-Soriano and Nielsen 2009), f in large sample sizes shows high variance.

Supplemental Figure S3



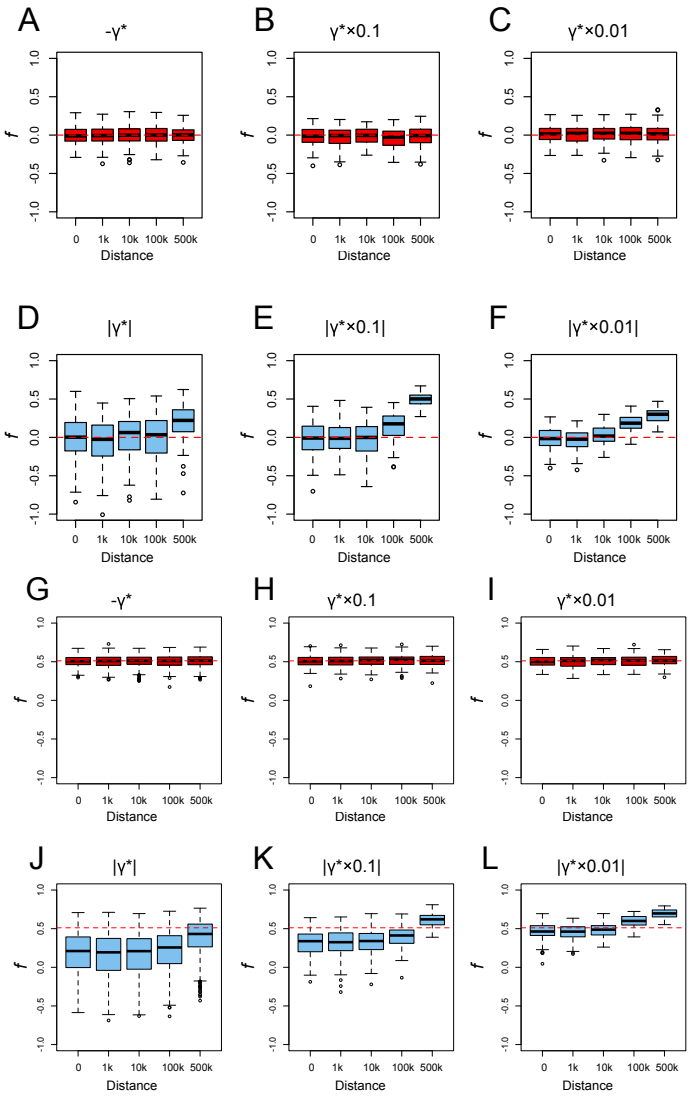
Supplemental Figure S3. Effect of recombination rate on estimates of *f*. Exons were classified into 20 groups based on the degree of the local recombination rate. The estimates of *f* were computed for nonsynonymous sites (A) and synonymous sites (B) in EA (red) and AA (blue). The slope of the dashed line was obtained from a linear regression model between average recombination rates and the estimates of *f*, showing that estimates of *f* are robust to recombination rate heterogeneity.

Supplemental Figure S4



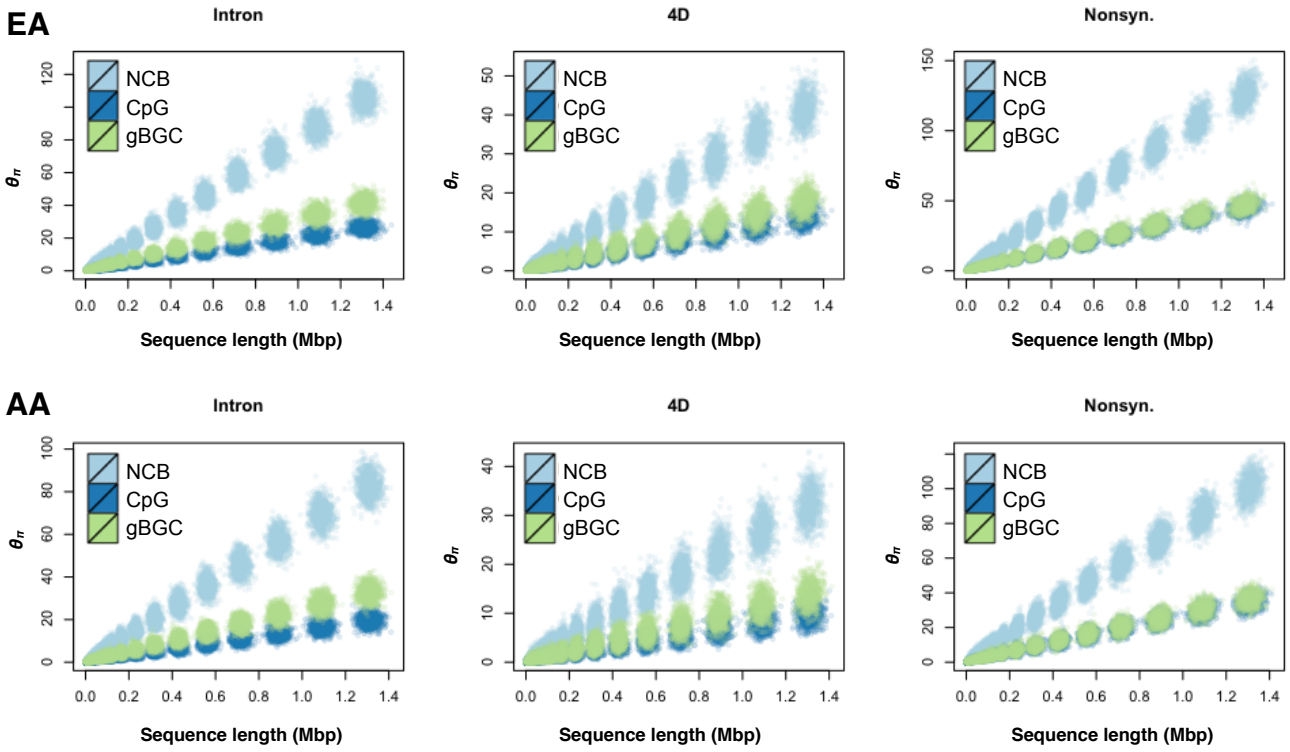
Supplemental Figure S4. Effect of sample size and selection coefficient on estimates of *f*. For a given selection coefficient, *f* was computed with varying sample size (2n = 10, 100,1000,10000) under a realistic demographic model of recent population expansion. For larger sample sizes, estimates of *f* captured more rare deleterious alleles for a fixed selection coefficient. γ* depicts the baseline model of selection, obtained from the Gamma distribution of selection coefficients for nonsynonymous sites estimated in humans (Boyko et al. 2008).

Supplemental Figure S5



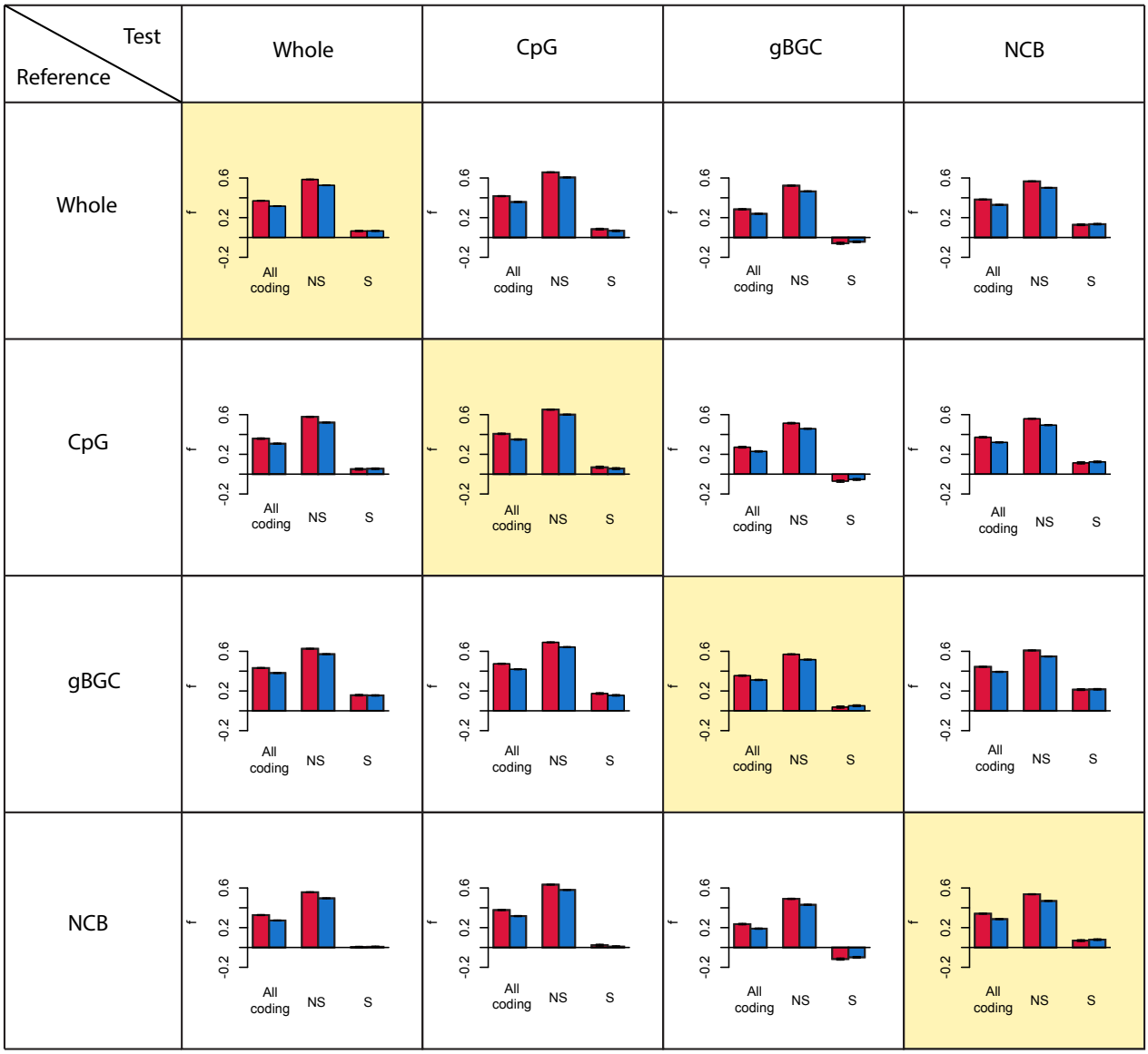
Supplemental Figure S5. Estimates of hitchhiking and background selection on estimates of *f*. With varying distances between reference sites (noncoding sequence) and coding region, *f* was estimated at neutral (A-F) and non-neutral sites ($\gamma * \times 0.1$;G-L) that were linked to nonsynonymous sites under purifying (A, B, C, G, H, I) and positive selection (D, E, F, J, K ,L) in the coding region. The dashed red line denotes the expected value of *f*. γ* depicts the baseline model of selection, obtained from the Gamma distribution of selection coefficients for nonsynonymous sites in human (Boyko et al. 2008). It is interesting to note that the hitchhiking effect in the cases of J, K, and L is analogous to the interaction term in the mixture of difference types of selections (Supplemental note 5.2, Figure S12).

Supplemental Figure S6



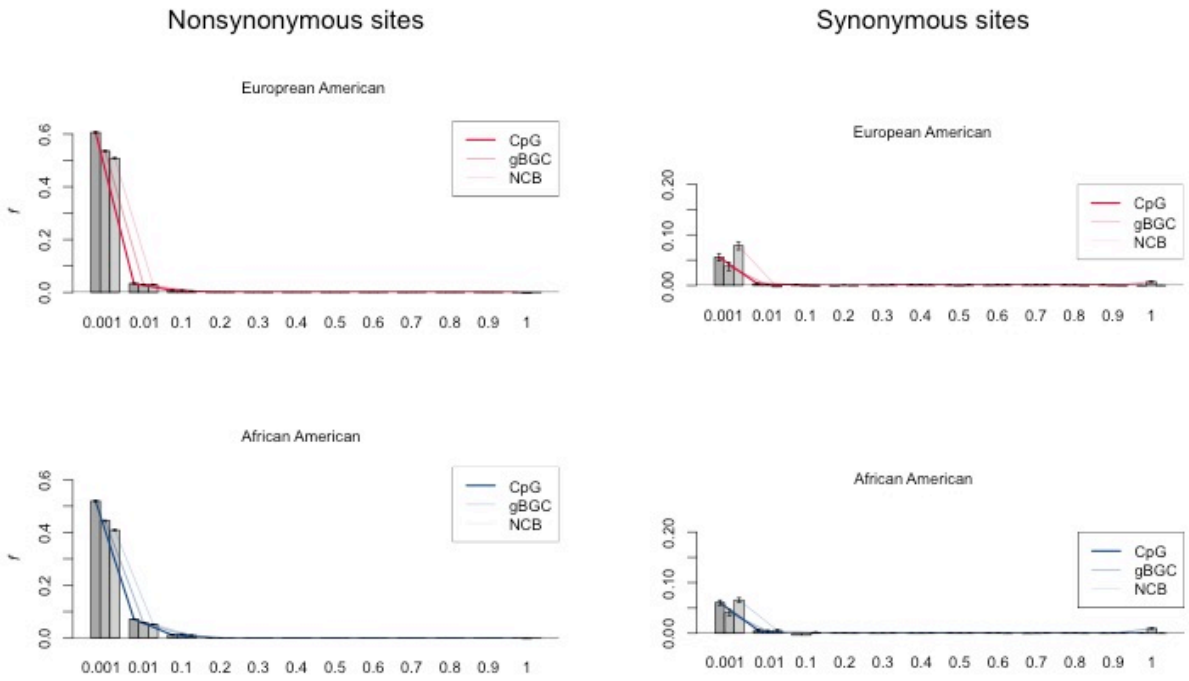
Supplemental Figure S6. Effect of sequence length on the estimator of θ. Give a number of genes, *n*, where *n* = 10, 20,...,100, the total length of coding sequences and nucleotide diversity was computed in each genomic contexts, including CpG, gBGC, and NCB classes, at intronic sites, synonymous sites, and nonsynonymous sites. There was a linear relationship between the amount of observable SNVs and sequence length, where the slopes were different between genomic contexts.

Supplemental Figure S7



Supplemental Figure S7. Effect of genomic context mismatch on *f*. For all combinations of genomic contexts for test and reference sites, average estimates of *f* were computed by bootstrapping 100 replicates of 1,000 randomly selected exons. Given its high rate of fixation,, the gBGC group reduced or inflated the estimates of *f* when used for test or reference sites, respectively, compared with estimates of *f* from the matching genomic context. Diagonal panels depict the estimates of *f* used in the study that controlled for the possible confounding effect of mutation rate heterogeneity in test and reference model.

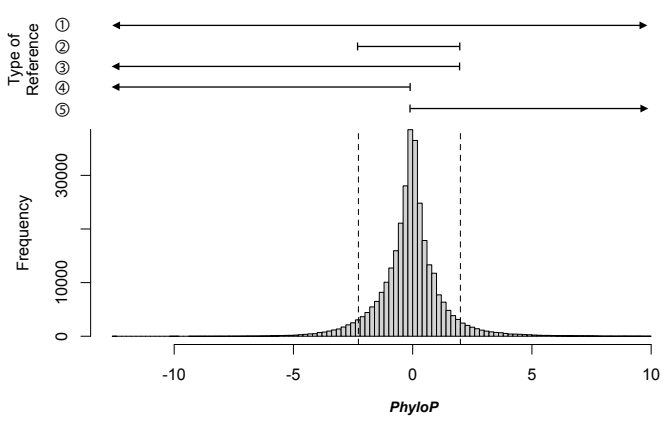
Supplemental Figure S8



Supplemental Figure S8. Effect of genomic context on the site frequency spectrum of non-neutral mutations. For rare (0.001) nonsynonymous mutations, the estimates of *f* were significantly (Wilcoxon-test *p*-value < 10^{-16}) high at CpG sites compared to gBGC and NCB sites.

Supplemental Figure S9

A

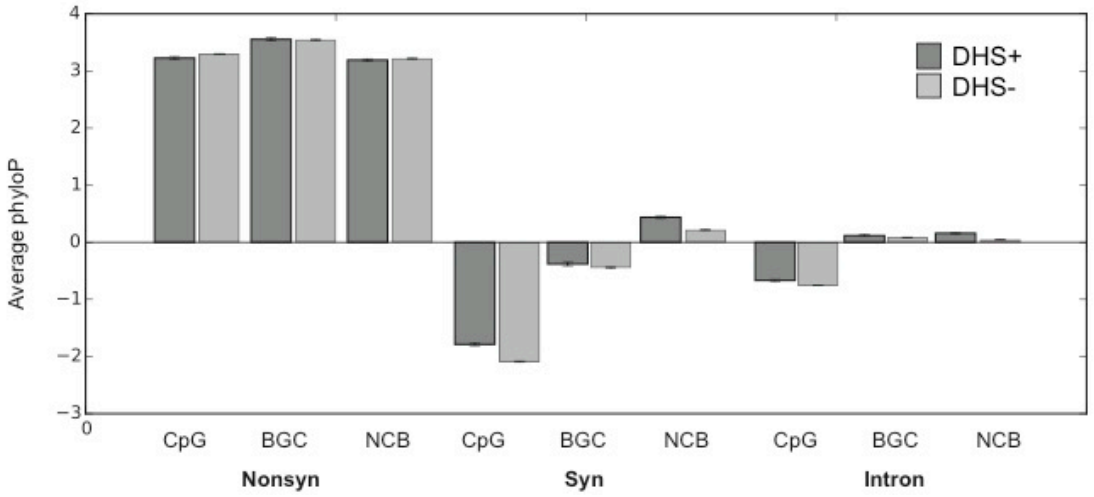


B

EA	Coding region	Nonsynonymous sites				synonymous sites			
	Type of reference	All	CpG	gBGC	NCB	All	CpG	gBGC	NCB
Ref. (Intronic sites)	① All	0.584787	0.656977	0.566285	0.534931	0.071963	0.091511	0.035003	0.073476
	② 90% Interquantile (-2.275<PhyloP<2.001)	0.587642	0.65247	0.574447	0.539411	0.078343	0.079575	0.053164	0.082401
	③ Remove conserved (5%) (PhyloP < 2.001)	0.593827	0.659933	0.579833	0.545405	0.092168	0.09934	0.065148	0.094342
	④ Remove all conserved (PhyloP < 0)	0.626407	0.674558	0.617294	0.588937	0.164987	0.138074	0.148495	0.181068
	⑤ Use only conserved(PhyloP > 0)	0.523265	0.601103	0.502432	0.467345	-0.061512	-0.053452	-0.096712	-0.057646

AA	Coding region	Nonsynonymous				synonymous			
	Type of reference	All	CpG	gBGC	NCB	All	CpG	gBGC	NCB
Ref. (Intronic sites)	① All	0.524501	0.606988	0.509178	0.464599	0.067464	0.075796	0.045592	0.072616
	② 90% Interquantile (-2.275<PhyloP<2.001)	0.527029	0.603035	0.516153	0.46916	0.072423	0.066501	0.059155	0.080516
	③ Remove conserved (5%) (PhyloP < 2.001)	0.532899	0.609475	0.522495	0.474067	0.083933	0.081646	0.071486	0.089014
	④ Remove all conserved (PhyloP < 0)	0.565573	0.624049	0.566768	0.512409	0.148013	0.115916	0.157577	0.155428
	⑤ Use only conserved(PhyloP > 0)	0.46438	0.552058	0.43899	0.40491	-0.048022	-0.050672	-0.083317	-0.029855

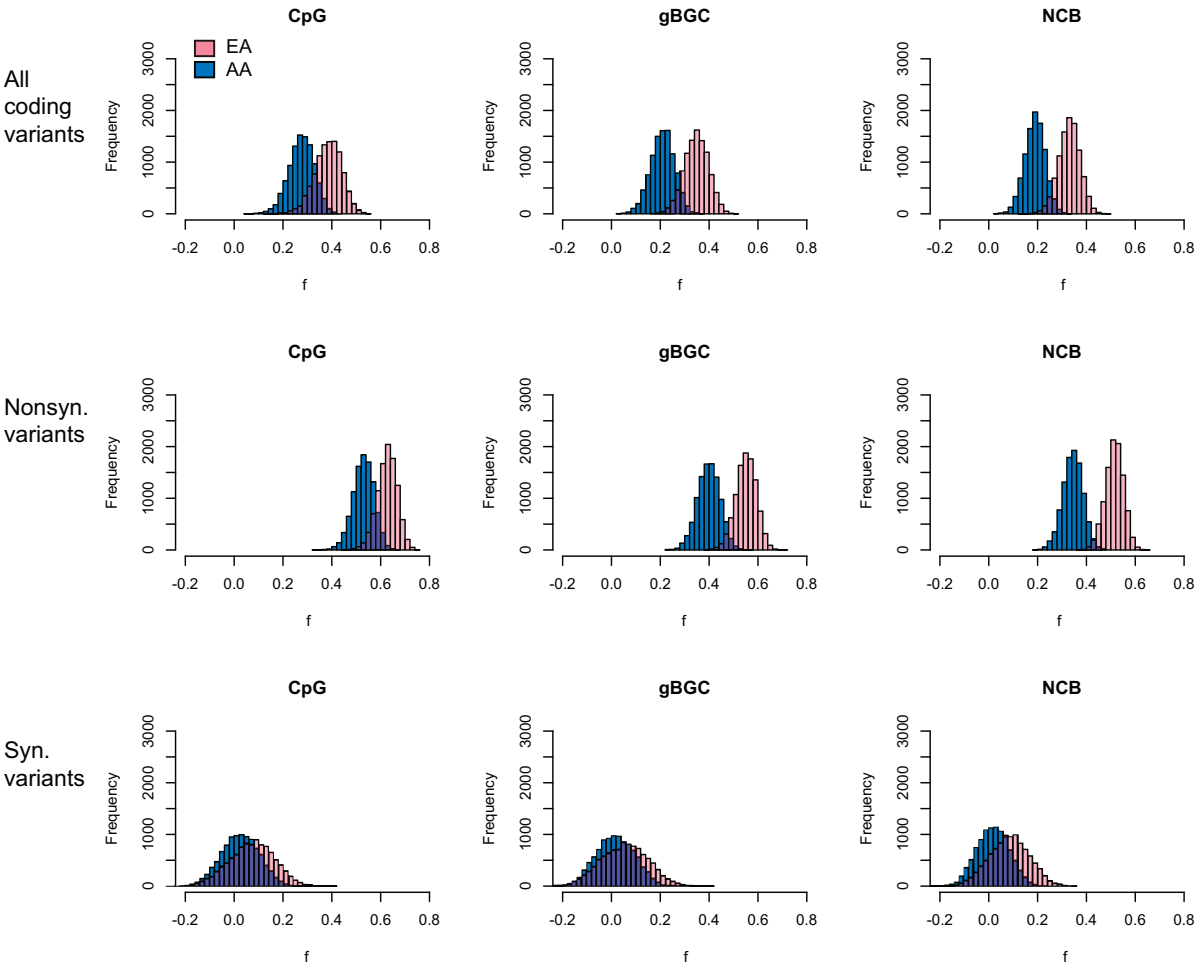
C



Supplemental Figure S9. Effect of conservation in reference model on the estimates of *f*. (A) distribution of PhyloP scores obtained from 100 vertebrates basewise conservation for intronic variants. The type of reference

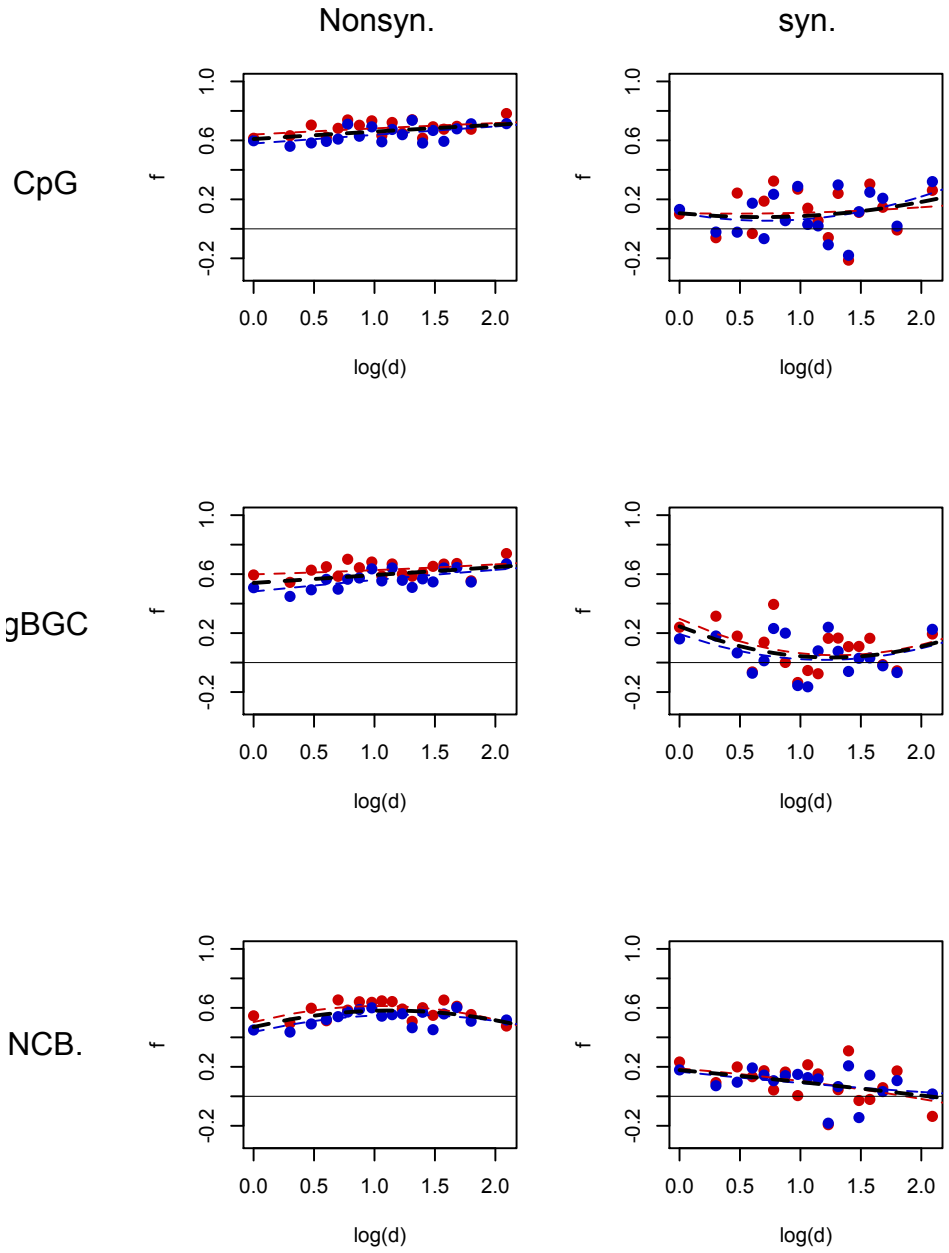
represents range of PhyloP scores for obtaining variants as a reference model. For example, the type (2), used in this study, includes 90% interquartile range of the distribution. In the other word, this excludes mutations assigned to top 5% of highly conserved sites as well as top 5% of highly accelerated sites. (B) Summary of the estimates of f in nonsynonymous and synonymous variants with using different type of conservation cutoff for the reference sites. As a final reference, type (2) was chosen to filtering out top 5% of highly conserved sites as well as top 5% of highly accelerated sites from intronic sites. (C) Average PhyloP scores of variants in each genomic contexts.

Supplemental Figure S10



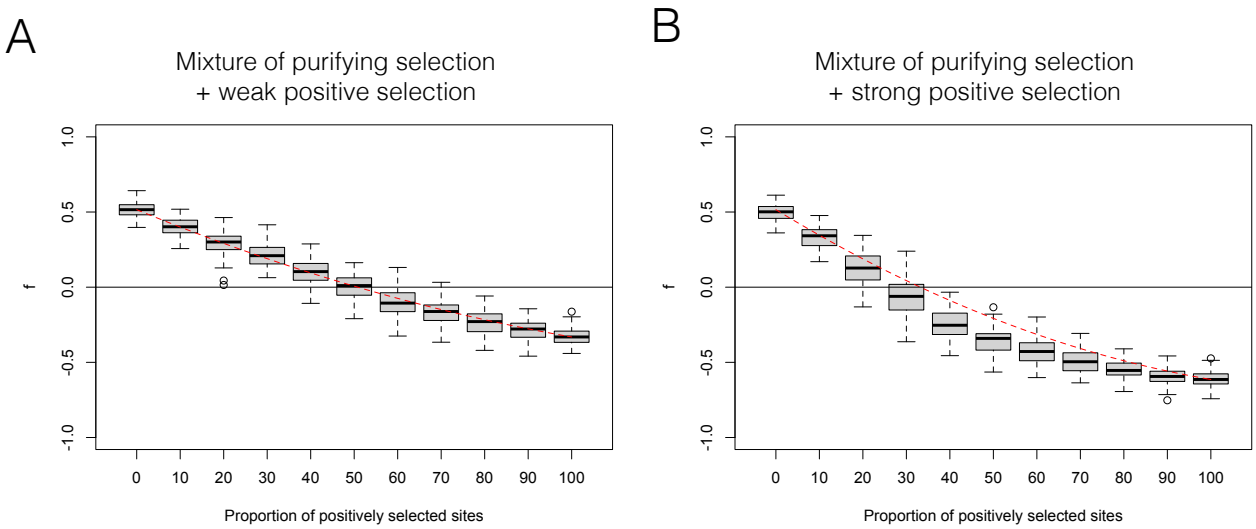
Supplemental Figure S10. Distribution of estimates of *f* in EA and that in AA. To compare average estimates of *f* in AA and EA, two bootstrapping was applied: first randomly selected *n* individuals from EA, *n* to be the number of individuals in AA (*n*= 2217), and than bootstrapped 1,000 genes with replacement to estimate the *f* for the sampled individual in EA and AA. All comparisons between EA and AA are statistically significant (Wilcoxon p $< 10^{-16}$).

Supplemental Figure S11



Supplemental Figure S11. The effect of genomic contexts and connectivity (*d*) of protein-protein interaction network on the estimates of *f*. For a group of genes having same degree of connectivity (*d*), variants of genes were separated into subgroups of genomic contexts in EA (red) and in AA (blue). Solid and dashed lines depict regression lines of combining EA and AA and each population, respectively.

Supplemental Figure S12



Supplemental Figure S12. Effect of mixture of deleterious and advantageous mutations on *f*. (A) mixture of sites under purifying selection and sites under weak positive selection, where selection coefficients were drawn from Gamma distribution with γ*/10 and |γ*|/100, respectively. (B) mixture of sites under purifying selection and sites under positive selection, where selection coefficients were drawn from Gamma distribution with -γ*/10 and |γ*|/10. The proportion of positively selection sites equals 1 – the proportion of negatively selection sites. Dashed line reflects expected average fraction of multiplicative effect of purifying selection and positive selection. For example, if the test sites consisted of a 80:20 mixture of γ*/10 (f_{Neg}=0.50) and |γ*|/10 (f_{Pos}= -0.61), then f is estimated to be $f_{Neg} + f_{Pos} + 2*f_{Neg}*f_{Pos} = 0.18$. γ* depicts the baseline model of selection, obtained from the Gamma distribution of selection coefficients for nonsynonymous sites in human (Boyko et al. 2008).

Supplemental References

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS genetics* **4**: e1000083.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**: 783-796.

Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. 2013. A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *PLoS genetics* **9**: e1003684.

Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome research* **15**: 901-913.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405-1413.

Fu YX, Li W-hH. 1993. Statistical tests of neutrality of mutations. *Genetics* **133**: 693-709.

Gazave E, Chang D, Clark AG, Keinan A. 2013. Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. *Genetics* **195**: 969-978.

Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science (New York, NY)* **336**: 740-743.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research* **20**: 110-121.

Ramirez-Soriano A, Nielsen R. 2009. Correcting estimators of theta and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process. *Genetics* **181**: 701-710.

Stephan W. 2010. Genetic hitchhiking versus background selection: the controversy and its implications. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* **365**: 1245-1253.

Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, Raubitschek A, Ziegler S, LeProust EM, Akey JM et al. 2013. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* **342**: 1367-1372.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437-460.

Tennessen Ja, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, NY)* **337**: 64-69.

Watterson G. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**: 256-276.

Ying H, Huttley G. 2011. Exploiting CpG hypermutability to identify phenotypically significant variation within human protein-coding genes. *Genome biology and evolution* **3**: 938-949.