Nonparametric inference of the distribution of fitness effects across functional categories in humans

Fernando Racimo[1,*], Joshua G. Schraiber[1],

1 Department of Integrative Biology, Univeristy of California, Berkeley, CA, USA

∗ E-mail: fernandoracimo@gmail.com

Abstract

Quantifying the proportion of polymorphic mutations that are deleterious or neutral is of fundamental importance to our understanding of evolution, disease genetics and the maintenance of variation genome-wide. Here, we develop an approximation to the distribution of fitness effects (DFE) of segregating single-nucleotide mutations in humans. Unlike previous methods, we do not assume that synonymous mutations are neutral, or rely on fitting the DFE of new nonsynonymous mutations to a particular parametric probability distribution, which is poorly motivated on a biological level. We rely on a previously developed method that utilizes a variety of published annotations (including conservation scores, protein deleteriousness estimates and regulatory data) to score all mutations in the human genome based on how likely they are to be affected by negative selection, controlling for mutation rate. We map this score to a scale of fitness coefficients via maximum likelihood using diffusion theory and a Poisson random field model. We then use our coefficient mapping to quantify the distribution of all scored single-nucleotide polymorphisms in Yoruba and Europeans. Our method serves to approximate the DFE of any type of segregating mutations, regardless of its genomic consequence, and so allows us to compare the proportion of mutations that are negatively selected or neutral across various genomic categories, including different types of regulatory sites. We observe that the distribution of intergenic polymorphisms is highly leptokurtic, with a strong peak at neutrality, while the distribution of nonsynonymous polymorphisms is bimodal, with a neutral peak and a second peak at $s \approx -10^{-4}$. Other types of polymorphisms have shapes that fall roughly in between these two.

Author Summary

The relative frequencies of polymorphic mutations that are deleterious, nearly neutral and neutral is traditionally called the distribution of fitness effects (DFE). Obtaining an accurate approximation to this distribution in humans can help us understand the nature of disease and the mechanisms by which variation is maintained in the genome. Previous methods to approximate this distribution have relied on fitting the DFE of new mutations to standard parametric probability distributions, like a normal or an exponential distribution. Here, we provide a novel method that does away with using parametric DFE approximations by relying on genomic scores designed to reflect the strength of negative selection operating on any site in the human genome. We use a maximum likelihood mapping approach to fit these scores to a scale of neutral and negative fitness coefficients. Finally, we compare the shape of the DFEs we obtain from this mapping across populations as well as different types of functional categories. We observe a highly leptokurtic distribution of polymorphisms, with a strong peak at neutrality, as well as a second peak of deleterious effects when restricting to nonsynonymous polymorphisms.

Introduction

Genetic variation within species is shaped by a variety of evolutionary processes, including mutation, demography, and natural selection. With the advent of whole-genome sequencing, we can make unprecedented inferences about these and other processes by analyzing population genomic data. An important goal is to understand the extent to which segregating genetic variants are impacted by natural selection, and to quantify the intensity of natural selection acting genome-wide. Understanding the prevalence of different modes of selection on a genomic scale has wide-ranging implications across evolutionary and medical genetics. For instance, genome-wide association studies (GWAS) are searching for mutations associated with disease in large samples of humans. Because mutations associated with disease are *a priori* likely to be deleterious, quantifying the portion of mutations that are deleterious along with their average effects could have significant implications for the design and interpretation of GWAS. Moreover, recently, the ENCODE project [1] has claimed that much of the genome is involved in some kind of vital molecular function. Although this has been disputed [2], quantifying the DFE in noncoding regions is a first step toward understanding the fitness implications of rampant functional activity at the genomic level.

Traditionally, studies have sought to estimate the distribution of fitness effects (DFE) for nonsynonymous mutations by using summary statistics based on the number of polymorphisms and substitutions [3–5] and/or the full frequency spectrum [6–8]. These studies typically assume that synonymous variation is neutral. Many of these analyses suggest that the distribution of deleterious fitness effects is strongly leptokurtic; that is, while most nonsynonymous mutations are nearly neutral, there is a significant probability that an amino acid changing mutation will be strongly deleterious. While these studies were limited to analysis of protein-coding genes, recently work has focused on quantifying the DFE in regulatory regions, including short interspersed genomic elements such as enhancers [9, 10] and cis-regulatory regions [11]. A review of many of these approaches can be found in ref. [12].

There are several obstacles to quantifying the DFE of new or segregating mutations genome-wide. First, inferences about the DFE are confounded by demography [13]. For example, a high proportion of low frequency derived alleles is a signature of negative selection, but can also be caused by recent population growth [14]. Hence, a well-supported demographic model must be used to appropriately control for population history when inferring the DFE. Second, most current methods rely on dividing up polymorphisms into either putatively neutral or putatively selected sites (for example, synonymous and nonsynonymous sites). Because of the reduced resolution afforded by having only two classes of sites, these studies have relied on fitting the DFE of new mutations to a parametric distribution, typically an exponential or gamma distribution [3, 7]. While flexible, these distributions may miss some important features of the DFE [15]. For example, mutation accumulation experiments suggest that the DFE may be bimodal, with most mutations either being nearly neutral or strongly deleterious, with very few in between [16, 17]. Thus, fitting a parametric distribution with a single mode may not capture all the relevant information about the DFE (but see [18] for an example of fitting a multimodal DFE to population genetic data and [15, 19] for nonparametric approaches to estimating the DFE of new amino-acid changing mutations). Finally, previous studies have been restricted to analyzing specific subclasses of mutations (e.g. nonsynonymous, enhancers, etc.) because until recently, no single metric existed that could serve to compare the disruptive potential of any type of variant, regardless of its genomic consequence.

Recently, Kircher et al. [20] developed a method to synthesize a large number of annotations into a single score to predict the pathogenicity or disruptive potential of any mutation in the genome. It is based on an analysis comparing real and simulated changes that occurred in the human lineage since the human-chimpanzee ancestor, and that are now fixed in present-day humans. The method relies on the realistic assumption that the set of real changes is depleted of deleterious variation due to the action of negative selection, which has pruned away disruptive variants, while the simulated set is not depleted of such variation. A support vector machine (SVM) was trained to distinguish the real from the simulated changes using a kernel of 63 annotations (including conservation scores, regulatory data and protein deleteriousness scores), and then used to assign a score (C-score) to all possible single-nucleotide changes in the human genome, controlling for local variation in mutation rates. These C-scores are meant to be predictors of how disruptive a given change may be, and are comparable across all types of sites (nonsynonymous, synonymous, regulatory, intronic or intergenic). Thus, they allow for a strict ranking of predicted functional disruption for mutations that may not be otherwise comparable. C-scores are PHRED scaled, with larger values corresponding to more disruptive effects.

Importantly, human-specific genetic variation patterns are not used as input to train the C-score SVM. In this work, we make use of the C-scores to provide a fine-grained stratification of deleteriousness in modern human populations. Using the 1000 Genomes dataset [21, 22], we take advantage of the Poisson random field model [23, 24] with a realistic model of human demographic history to fit a maximum likelihood selection coefficient for each C-score, creating a mapping from C-scores to selection coefficients. Using this mapping, we obtain a high-resolution picture of the DFE in Europeans and Africans, and explore the DFE of different mutational consequences.

Results

A mapping from C-scores to selection coefficients

To map C-scores to selective coefficients, we obtained allele frequency information from 176 low-coverage Yoruba (YRI) chromosomes from the 1000 Genomes Project Phase 1 data [21, 22]. We tested only models of neutral evolution and negative selection, because C-scores are uninformative about adaptive vs. deleterious disruption (i.e. a high C-score could either reflect a highly deleterious change or a highly adaptive change), and, because we are using polymorphism data only, positive selection should contribute little to the site-frequency spectrum [25].

We began by binning sites into C-scores rounded up to the nearest integer and computed the site frequency spectrum for each bin (Figure S1). We then fit the lowest possible C-score ($C = 0$), presumed to be neutral, to different models of demographic history. We compared constant population size, exponential growth (fitting the parameters by maximum likelihood; see Methods) and the model inferred in Harris and Nielsen [26] from the distribution of tracts of identity by state (IBS) (Figure S2). We find that the constant population size and the Harris and Nielsen models fit the data approximately equally well and better than any of the exponential growth models we tried. We picked the Harris and Nielsen model for downstream analyses, as it is based on haplotype information (the distribution of tracts of identity by state), and may thus be a better reflection of the true demographic history.

We next fit a selection coefficient to the site frequency spectrum for each $C < 40$ using maximum likelihood (see Methods). We restricted to $C < 40$ because very few sites have $C \geq 40$, and hence estimates of the selection coefficients for those C-scores are unreliable. Predictably, the lowest C-score bin ($C = 0$) fits the neutral model ($s = 0$) best, as that was the bin used in the neutral demographic fitting. In addition, the next highest bin ($C = 1$) also maps to s=0. Figure S3 shows that the site frequency spectra of the C-score bins are well-modeled by our maximum likelihood fits.

We aimed to test the robustness of the selection coefficient estimates within each bin. We were specifically concerned about highly deleterious bins, which are composed of a smaller number of segregating sites than neutral or nearly neutral bins, and could produce unstable or biased estimates. We obtained bootstrapped confidence intervals for each bin and observe that the mappings are relatively stable up to $C = 36$. As expected, the standard deviation of the bootstrap estimates is strongly negatively correlated with the sample-size per bin (Figure S4, Pearson correlation coefficient = -0.933). Thus, most of the increase in the width of the confidence intervals observed at higher C-score bins can be explained by the small number of polymorphisms available in those bins, and is likely not the result of other unaccounted processes, such as positive selection, operating exclusively on highly scored polymorphisms.

After removing the C-score bins that best fit the neutral model, the remaining C-scores plotted as a function of $\log_{10}(-s)$ appear to have an odd-degree polynomial shape. Using least-squares regression, we fit different polynomial functions to the mapping, as well as an inverted logistic curve, to obtain a continuous function from C-scores to $\log_1 0(-s)$. Although the 5th and 7th degree polynomials fit approximately equally well (residuals of .1962 and .1819, respectively), we chose the 5th degree fit because the 7th degree mapping showed signs of overfitting (Figure S5). Figure 1 shows our mapping of C-scores to selection coefficients, including confidence intervals obtained by bootstrapping the data in each bin 100 times. Interestingly, there is a plateau from approximately $C = 10$ to $C = 30$ where a variety of C

scores correspond to identical selection coefficients. After approximately $C = 30$, the strength of selection increases substantially.

To test for the robustness of our mapping, we performed the same fitting procedure on a variety of other conservation and deleteriousness scores (see Methods). Figure S6 shows that mappings are fairly consistent across different choices of scores, except for highly deleterious bins, which we were already excluding from the analysis. In the following, we only report results using the C-score mapping, as this score has been shown to be a better correlate to functional disruption and pathogenicity than all the other conservation scores mentioned above, and also controls for mutation rate variation across the genome, while other scores do not [20]. Additionally, Figure S7 show that this score is the best at distinguishing nonsynonymous from synonymous changes.

The distribution of fitness effects of segregating mutations in Yorubans and Europeans

Using the C-score-to-selection coefficient mapping, we obtained the DFE of segregating polymorphisms in Yoruba individuals. This distribution is highly leptokurtic when all polymorphisms are considered (Figure 2, black dashed line), with a considerably high peak at neutrality and a long tail of deleterious mutations, as has been observed before when estimating the DFE of coding sequences [3, 5–7, 13]. Interestingly, we observe a pronounced drop in frequency for values of $s < -10^{-4}$. We note that this is not due to our capping our mapping at $C = 39$ as the selection coefficients we are able to map are of a greater magnitude than this drop.

When we partition the data by the genomic consequence of the polymorphisms, some classes exhibit a peak of highly deleterious changes around $s = -10^{-4}$. This peak results in a bimodal distribution that is especially pronounced for nonsynonymous sites (Figure 2, red line), and is almost non-existent for intergenic sites (Figure 2, pink line). Synonymous polymorphisms also show a highly deleterious peak; this may indicate selection for optimal codon usage [27] and may be consistent with a recent finding of strong synonymous selection in *Drosophila* [28], but could also result from widespread patterns of background selection during human evolution [35,36]. Other types of polymorphisms—like splice site, 3' UTR, 5' UTR and regulatory mutations—have a bimodal distribution, though with an smaller deleterious peaks than for coding sites (Figure 2). We can compare the selection coefficient distributions to the distributions of unmapped C-scores (Figure S8) which are much less tightly peaked at intermediate deleterious values and do not show a sharp decrease in density for highly deleterious polymorphisms, as does the s distribution in Figure 2. We show various statistics calculated on each of the selection coefficient distributions in Table 1.

Next, we partitioned the data by whether the polymorphisms were found in the GWAS database [29] or not (Figure S9, Table 1). We observe a second deleterious peak among the GWAS SNPs, too, but these SNPs are also highly enriched for neutral polymorphisms. In addition, we classified polymorphisms by different ENCODE categories using the RegulomeDB classifier [30] (Figure S10, Table 2).

Finally, we compared the distribution of fitness effects between Yoruba and Europeans. We observe a slight excess of deleterious sites in Europeans, consistent with previous studies [6,31] (Figure 3). This is especially prominent for nonsynonymous polymorphisms with $s < -10^{-4}$: we estimate that 3.1% of nonsynonymous segregating polymorphisms in Europeans fall in this category, while the same is true for 2.5% of nonsynonymous segregating polymorphims in Yoruba. However, we caution that this is based on inferring the C-to-s mapping at values of s for which there exist very few segregating mutations.

Discussion

The distribution of fitness effects (DFE) describes the proportion of mutations with given selection coefficients. Knowledge of the DFE has profound implications for our understanding of evolution and health. We believe ours is the first study to estimate the distribution of deleterious fitness effects in human polymorphisms genome-wide, without assuming a parametric probability distribution for the DFE. We infer a highly leptokurtic distribution for all polymorphisms, with a sudden drop in density at $s \approx -10^{-4}$, which may be the cutoff between weakly deleterious and nearly neutral segregating mutations and highly deleterious mutations that are easily pruned away by negative selection.

Our inferred non-synonymous distribution is bimodal and looks very similar to the one obtained for nonsynonymous mutations in Drosophila in ref. [5], with a peak at neutrality and another peak at $s \approx 0.9 \times 10^{-4}$, albeit with the difference that the neutral peak we observe in humans is relatively larger. Several experimental studies have also shown that non-synonymous non-lethal mutations tend to have a multimodal DFE in model organisms [32,33] (see ref. [12] for a comprehensive review). We note that it is impossible to obtain such kinds of distributions using a gamma or lognormal probability distribution unless one approximates bimodality by assuming a second, separate class of nonsynonymous mutations that are completely neutral and do not follow the best-fitting probability distribution [5, 7, 13, 18].

Importantly, unlike previous studies, we also obtain DFEs for other types of mutations, including synonymous, splice site, 3' UTR, 5' UTR and regulatory polymorphisms, which exhibit bimodality to a lesser degree than the nonsynonymous DFE. In particular, 5' UTR changes constitute the category with the smallest proportion of neutral polymorphisms after nonsynonymous changes, likely reflecting selection on gene regulation upstream of coding sequences. Futhermore, distributions corresponding to mutations in UTR and regulatory regions have a less pronounced trough between the two peaks than the ones observed among coding mutations, suggesting that the magnitude of deleterious effects is more uniformly distributed in non-coding regions. In contrast, missense mutations appear to have more of an "all-or-nothing" effect, as would perhaps be expected when replacing an amino acid inside a protein.

Our method does not assume that synonymous changes are neutral, as do other studies [3, 5, 13]. Given that there is evidence for selection for codon usage in humans [34] and that our inferred DFE for synonymous polymorphisms also exhibits a highly-deleterious peak, the assumption that synonymous sites are neutral may no longer be viable. A second possibility is widespread patterns of background selection in human evolution [35, 36]. This could also lead to a depletion of synonymous mutations from the list of fixed human-chimpanzee differences, resulting in the SVM machine associating synonymous mutations with higher C-scores than one would expect under a model with no linked selection. In contrast, it seems intergenic polymorphisms are the class of sites most likely to be governed by neutrality. Because this class is so abundant, most of the signal observed when all polymorphisms are pooled together closely reflects the distribution observed for intergenic polymorphisms.

Our results have implications for GWAS, as we find a high proportion of GWAS SNPs to be neutral or nearly neutral, which could suggest a high rate of false positives in this type of association studies, although GWAS studies only aim to find polymorphisms linked to causative variants. Alternatively, if the effect size of many GWAS SNPs are sufficiently small, it is possible that many of them are not subject to strong selection.

Additionally, by stratifying our results based on different ENCODE categories, we can elucidate the fitness consequences of the molecular activity detected by ENCODE. We find the category with the lowest proportion of neutral polymorphisms to be the one corresponding to sites that have eQTL evidence as well as evidence for transcription factor (TF) binding, a matched TF motif, a matched DNase footprint and that are located in a DNase peak. In general, categories that combine many regulatory signals tend to show lower proportions of neutral mutations than those that do not, suggesting that data integration across distinct approaches to detecting selection and functionality is likely to do better than any individual approach [37]. Moreover, this suggests that much of the molecular activity detected by ENCODE may not have significant fitness consequences.

There are several limitations to our method. First, we have restricted ourselves to estimating the DFE of segregating mutations that have reached appreciable frequencies in the population. An extension of this approach would be to infer the DFE of new mutations from the DFE of segregating mutations genome-wide. Second, we assumed no dominance or epistasis. Future studies could attempt to incorporate a distribution of heterozygous and epistatic effects into our approach. In addition, we have assumed sites are independent and have therefore ignored the covariance between linked sites, which likely leads to an underestimatation of confidence intervals obtained from the bootstrapping. The free-recombination assumption may also affect inference due to Hill-Robertson interference between mutations subject to selection [38] as well as linked background selection affecting the SFS of neutral sites in the human genome [36]. This may be a more important issue in our case than other genic-only approaches because we are also including intergenic mutations in our analysis, so the space between analyzed polymorphisms is on average smaller than if we were only looking at coding polymorphisms [13]. We also assume no positive selection. This, however, should not be a major problem, because we are only basing our inferences on polymorphic sites and advantageous mutations contribute little to polymorphism, assuming $N_e s > 25$ [25]. One final limitation is that the type of inference performed here is only possible in species in which C-scores have been estimated (for now, humans only). Nevertheless, it should not be hard to obtain C-scores for other organisms in the future, although limitations on available annotations for non-human organisms may make the approximation to the fitness distribution less accurate.

Materials and Methods

Site frequency spectrum likelihoods

We used the theory developed by Evans *et al.* [39] to obtain the expected population site frequency spectrum with non-equilibrium demography. Writing $f(x,t)$ for the frequency spectrum at frequency x and time t and $g(x,t) := x(1-x)f(x,t)$, we can approximate the dynamics of $g(x,t)$ with selection and mutation by solving the following partial differential equation:

$$\frac{\partial}{\partial t}g(x,t) = -Sx(1-x)\frac{\partial}{\partial x}[g(x,t)] + \frac{x(1-x)}{2\rho(t)}\frac{\partial^2}{\partial x^2}[g(x,t)] \tag{1}$$

subject to boundary condition:

$$\lim_{x\downarrow 0} g(x,t) = \theta\rho(t) \tag{2}$$

where S is the population-scaled selection coefficient ($S = 2N(0)s$), θ is the population-scaled mutation rate ($\theta = 4N(0)\mu$) and $\rho(t) = N(t)/N(0)$ is the population size at time t relative to the population size at time 0. For the constant population size model, $\rho(t) = 1$, for the exponential growth model $\rho(t) = exp(Rt)$ where $R = 2N(0)r$ is the population scaled growth rate and for the model of Harris and Nielsen, $\rho(t)$ is piecewise defined according to their Figure 7.

We solve for $g(x,t)$ numerically in Mathematica, and can then compute the expected number of segregating sites with i copies of the derived allele out of a sample of n genes,

$$f_{n,i}(t) = \int_0^1 x^{i-1}(1-x)^{n-i-1}g(x,t)dx. \tag{3}$$

To compute the likelihood of the observed site frequency spectrum, $S = (s_1, s_2, \ldots s_{n-1})$ where s_i is the number of sites with i copies of the derived allele, for a given model, M, which includes both demography and selection, we observe that the probability that a given site in a sample of size n has i copies of the derived allele is

$$p_{n,i}(t) = \frac{f_{n,i}(t)}{\sum_{j=1}^{n-1} f_{n,j}(t)}. \tag{4}$$

Thus, the likelihood of S is

$$L(S|M) = \prod_{i=1}^{n-1} p_{n,i}. \tag{5}$$

We provide Mathematica scripts implementing this computation upon request.

Maximum likelihood fitting of exponential growth

The exponential growth model has two free parameters, r, the per generation growth rate and t, the total time of exponential growth. We first obtained the site frequency spectrum for all sites with $C = 0$. Next we solved $g(x,t)$ for the exponential growth model across a grid of t and r, and computed the likelihood of the data under each model.

Maximum likelihood fitting of selection coefficients

To find the maximum likelihood estimate of s for each C-score bin, we first obtained the site frequency spectrum corresponding to each C-score bin. Next, we solved $g(x, t)$ under the Harris and Nielsen demography for $\log_{10}(-s) \in [-6, -1.5]$ in steps of 0.05, along with $s = 0$. The selection coefficient with the highest likelihood was assigned to that C-score bin. After this assignment, the distributions were plotted using kernel density estimation with smoothing bandwith = 0.00001.

Testing robustness of the mapping

To test how robust the mapping of C-scores to selection coefficients is to different types of conservation scores, we obtained PhyloP [40] and PhastCons [41] scores derived from vertebrate, mammal and primate alignments (excluding humans), as well as GERP S scores [42], for all YRI SNPs. We attempted to equalize the range of all scores by PHRED-scaling them, i.e. converting each score to -$\log_{10}(p)$ where p is the probability of observing a change as or more disruptive / conserved (based on that particular score scale) among all polymorphic YRI sites. We note that this is different from the natural PHRED scale of C-scores (where p is the the probability of observing a score as or more disruptive among all possible, but not necessarily realized, mutations in the human genome), and so we also re-scaled the C-scores to produce a fair comparison. Then, we repeated the maximum likelihood mapping for each PHRED-scaled score in bins of 0.25 units (e.g. 0-0.125, 0.125-0.375, 0.375-0.625, etc).

Acknowledgments

We thank Montgomery Slatkin and Rasmus Nielsen for helpful advice and discussions, and Martin Kircher for providing the C-score data and other annotations.

References

1. Dunham I, Kundaje A, Aldred S, Collins P, Davis C, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

2. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, et al. (2013) On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome biology and evolution 5: 578–590.

3. Piganeau G, Eyre-Walker A (2003) Estimating the distribution of fitness effects from DNA sequence data: Implications for the molecular clock. Proceedings of the National Academy of Sciences 100: 10335–10340.

4. Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL (2003) Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. Journal of molecular evolution 57: S154–S164.

5. Loewe L, Charlesworth B, Bartolomé C, Nöel V (2006) Estimating selection on nonsynonymous mutations. Genetics 172: 1079–1092.

6. Keightley PD, Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177: 2251–2261.

7. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS genetics 4: e1000083.

8. Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M (2011) A population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS genetics 7: e1002395.

9. Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, et al. (2013) Genome-wide inference of natural selection on human transcription factor binding sites. Nature genetics .

10. Gronau I, Arbiza L, Mohammed J, Siepel A (2013) Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. Molecular biology and evolution 30: 1159–1171.

11. Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X, et al. (2009) Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS genetics 5: e1000592.

12. Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. Nature Reviews Genetics 8: 610–618.

13. Eyre-Walker A, Woolfit M, Phelps T (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173: 891–900.

14. Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, et al. (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proceedings of the National Academy of Sciences 102: 7882–7887.

15. Kousathanas A, Keightley PD (2013) A comparison of models to infer the distribution of fitness effects of new mutations. Genetics 193: 1197–1208.

16. Wloch DM, Szafraniec K, Borts RH, Korona R (2001) Direct estimate of the mutation rate and the distribution of fitness effects in the yeast Saccharomyces cerevisiae. Genetics 159: 441–452.

17. Sanjuán R, Moya A, Elena SF (2004) The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proceedings of the National Academy of Sciences of the United States of America 101: 8396–8401.

18. Loewe L, Charlesworth B (2006) Inferring the distribution of mutational effects on fitness in Drosophila. Biology Letters 2: 426–430.

19. Keightley PD, Eyre-Walker A (2010) What can we learn about the distribution of fitness effects of new mutations from dna sequence data? Philosophical Transactions of the Royal Society B: Biological Sciences 365: 1187–1193.

20. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, et al. (In press) A general framework for estimating the relative pathogenicity of human genetic variants .

21. Consortium GP (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.

22. Consortium GP (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

23. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132: 1161–1176.

24. Bustamante C, Nielsen R, Sawyer S, Olsen K, Purugganan M, et al. (2002) The cost of inbreeding in Arabidopsis. Nature 416: 531.

25. Smith NG, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415: 1022–1024.

26. Harris K, Nielsen R (2013) Inferring demographic history from a spectrum of shared haplotype lengths. PLoS genetics 9: e1003521.

27. Loewe L, Charlesworth B (2007) Background selection in single genes may explain patterns of codon bias. Genetics 175: 1381–1393.

28. Lawrie DS, Messer PW, Hershberg R, Petrov DA (2013) Strong Purifying Selection at Synonymous Sites in D. melanogaster. PLoS genetics 9: e1003527.

29. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences 106: 9362–9367.

30. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, et al. (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome research 22: 1790–1797.

31. Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, et al. (2008) Proportionally more deleterious genetic variation in European than in African populations. Nature 451: 994–997.

32. Davies EK, Peters AD, Keightley PD (1999) High frequency of cryptic deleterious mutations in Caenorhabditis elegans. Science 285: 1748–1751.

33. Keightley PD (1996) Nature of deleterious mutation load in Drosophila. Genetics 144: 1993–1999.

34. Plotkin JB, Robins H, Levine AJ (2004) Tissue-specific codon usage and the expression of human genes. Proceedings of the National Academy of Sciences of the United States of America 101: 12588–12591.

35. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, et al. (2011) Classic selective sweeps were rare in recent human evolution. Science 331: 920–924.

36. McVicker G, Gordon D, Davis C, Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. PLoS genetics 5: e1000471.

37. Scheinfeldt LB, Tishkoff SA (2013) Recent human adaptation: genomic approaches, interpretation and insights. Nature Reviews Genetics 14: 692–702.

38. McVean GA, Charlesworth B (2000) The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics 155: 929–944.

39. Evans SN, Shvets Y, Slatkin M (2007) Non-equilibrium theory of the allele frequency spectrum. Theoretical population biology 71: 109–119.

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

41. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research 15: 1034–1050.

42. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248–249.

390 # Figures

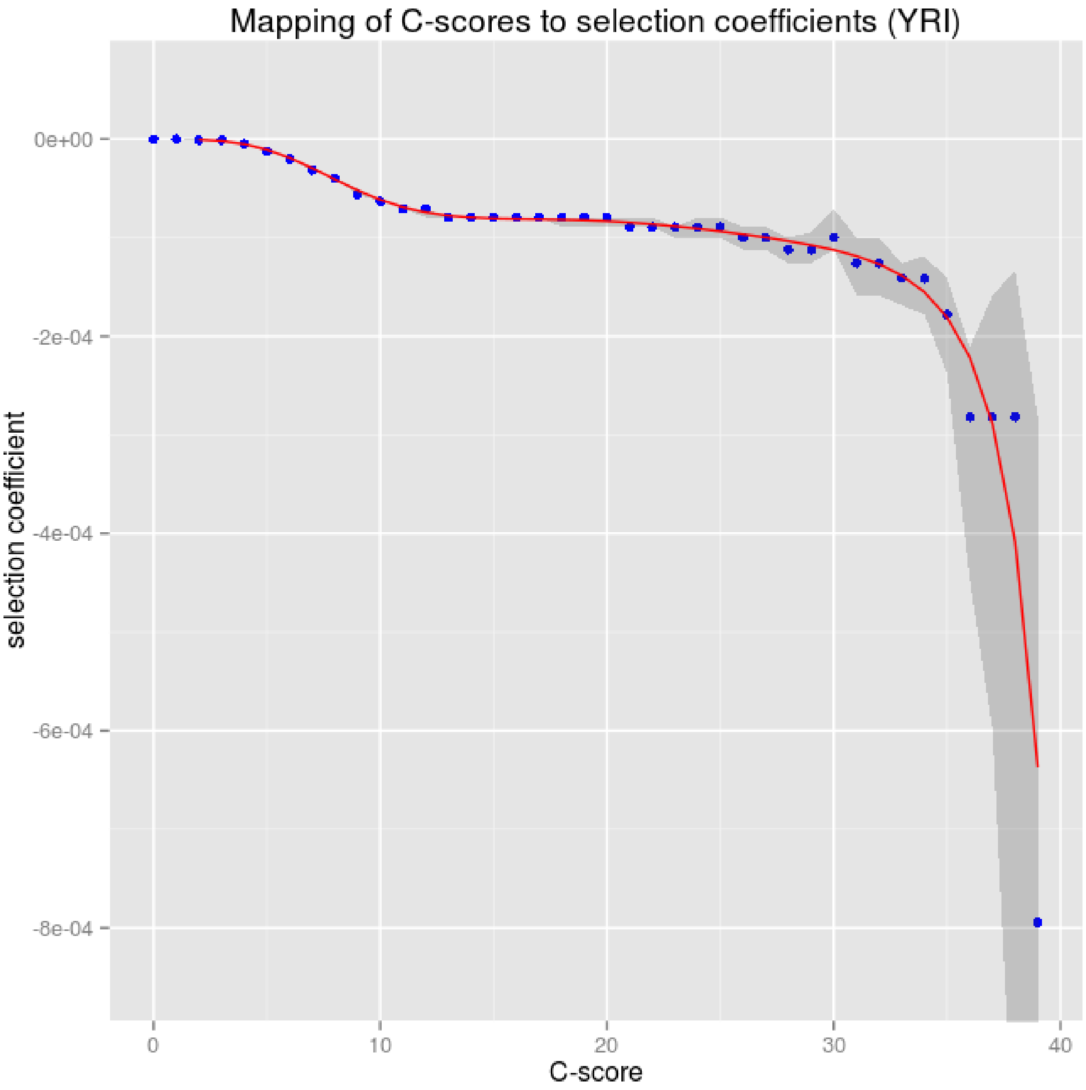

Figure 1. Mapping of C-scores to selection coefficients using YRI 1000G polymorphisms. Red dots represent the maximum likelihood selection coefficient corresponding to each C-score bin. The blue line is a polynomial fitted to the discrete mappings using partial least-squares regression on the mapping of C-scores to log-scaled selection coefficients (after excluding the neutral bins). The grey shade is a 95% confidence interval obtained from bootstrapping the data 100 times in each bin.

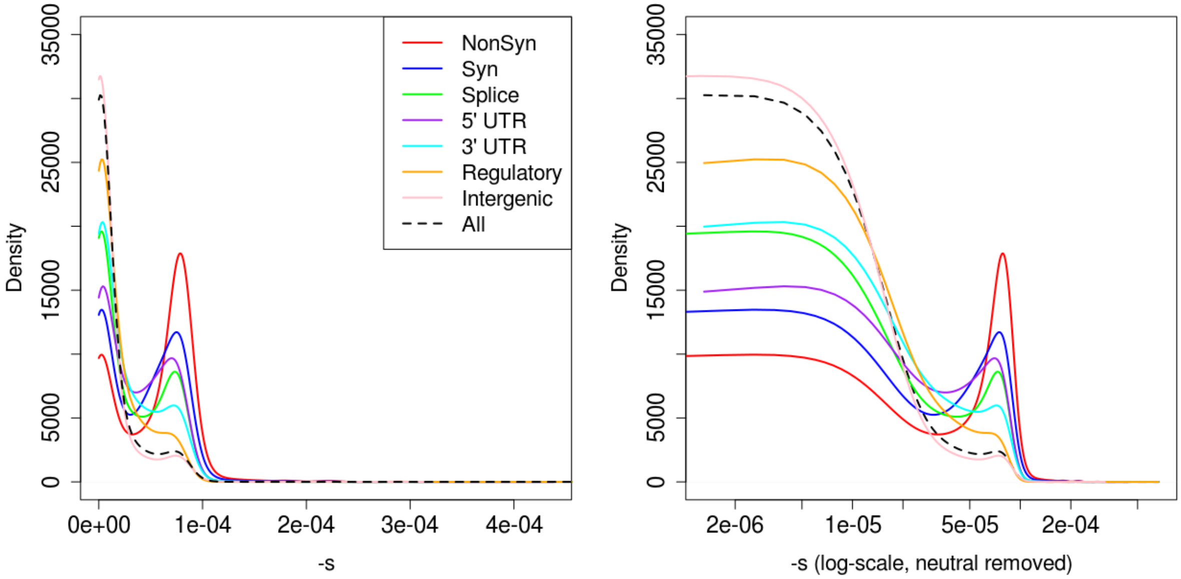

Figure 2. Distribution of fitness effects among YRI polymorphisms, partitioned by the genomic consequence of the mutated site. The right panel shows a zoomed-in version of the same distributions after removing neutral polymorphisms and log-scaling the x-axis. Consequences were determined using the Ensembl Variant Effect Predictor (v.2.5). If more than one consequence existed for a given SNP, that SNP was assigned to the most severe of the predicted categories, following the VEP's hierarchy of consequences. NonSyn = nonsynonymous. Syn = synonymous. Splice = splice site.

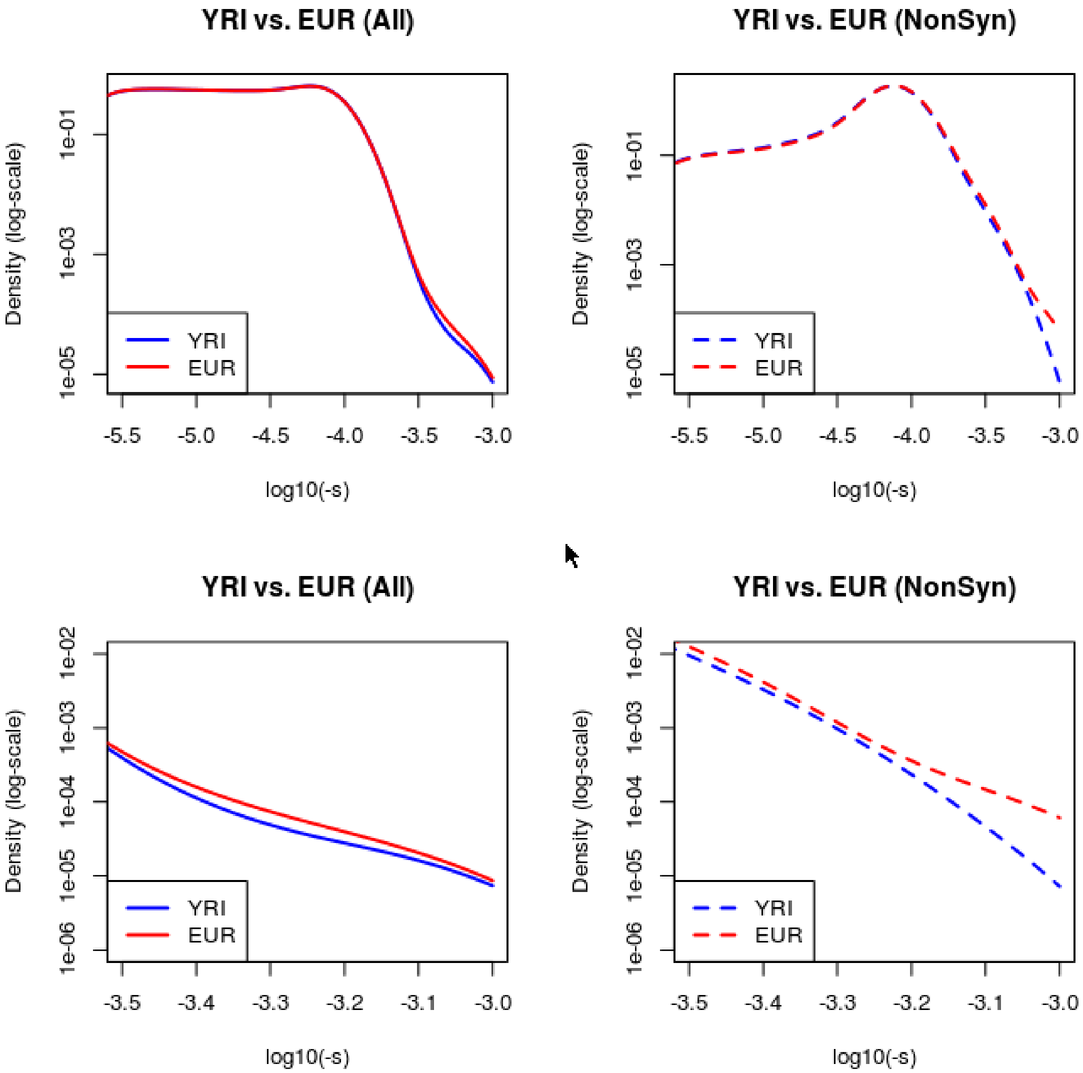

Figure 3. Comparison between the fitness effect densities corresponding to Yoruba and European polymorphisms with s $<$ 0. In all panels, the y-axis is on a log-scale. The density was computed using a smoothing bandwidth = 0.15. Left panels: distributions of all polymorphisms. Right panels: distributions of nonsynonymous polymorphisms. The bottom panels are a zoomed-in version of the top panels, focusing on highly deleterious mutations ($-3.5 < \log_{10}(-s) < -3$).

Tables

Table 1. Characteristics of fitness effect distributions estimated for YRI SNPs classified by different genomic consequence categories.

Category	Number of polymorphisms	Proportion s=0	Proportion $\lvert s\rvert > 10^{-5}$	Proportion $\lvert s\rvert > 5 * 10^{-5}$	Proportion $\lvert s\rvert > 10^{-4}$	Mean $\log_{10}(-s)$, for all $s \neq 0$	SD $\log_{10}(-s)$, for all $s \neq 0$
All	15956570	57.76%	25.26%	9.16%	0.08%	-4.8276	0.5054
Nonsynonymous	71242	17.80%	75.89%	60.95%	2.51%	-4.2885	0.3661
Synonymous	133797	23.05%	67.70%	46.45%	1.35%	-4.3989	0.4122
Splice site	21353	33.15%	52.98%	30.27%	0.30%	-4.5505	0.4796
5' UTR	53130	21.77%	65.17%	36.07%	0.30%	-4.5116	0.4396
3' UTR	169336	31.14%	52.71%	23.34%	0.37%	-4.6233	0.4681
Regulatory	1511150	39.18%	40.12%	13.45%	0.03%	-4.7734	0.4792
Intergenic	6211005	62.43%	21.33%	7.93%	0.12%	-4.8560	0.5140
GWAS	9673	45.90%	33.31%	12.28%	0.14%	-4.8028	0.5008

Table 2. Characteristics of fitness effect distributions estimated for YRI SNPs classified by different RegulomeDB regulatory categories.

RegulomeDB Category	Number of polymorphisms	Proportion s=0	Proportion $\lvert s\rvert > 10^{-5}$	Proportion $\lvert s\rvert > 5 * 10^{-5}$	Proportion $\lvert s\rvert > 10^{-4}$	Mean $\log_{10}(-s)$, for all $s \neq 0$	SD $\log_{10}(-s)$, for all $s \neq 0$
eQTL+TF binding+matched TF motif+matched DNase Footprint+DNase peak	274	28.83%	50.00%	20.80%	0.00%	-4.7019	0.4819
TF binding+matched TF motif+matched DNase Footprint+DNase peak	18080	32.09%	48.74%	22.02%	0.06%	-4.6733	0.4891
eQTL+TF binding+any motif+DNase Footprint+DNase peak	2140	32.71%	45.79%	16.07%	0.14%	-4.7487	0.4772
TF binding+any motif+DNase Footprint+DNase peak	174285	38.08%	41.94%	16.05%	0.08%	-4.7398	0.4893
eQTL+TF binding+any motif+DNase peak	1385	40.36%	40.43%	14.73%	0.14%	-4.7572	0.4923
TF binding+DNase peak	592313	40.53%	39.11%	13.84%	0.08%	-4.7690	0.4874
eQTL+TF binding+matched TF motif	46	43.48%	36.96%	21.74%	0.00%	-4.7214	0.5677
eQTL+TF binding / DNase peak	28697	44.88%	34.67%	11.88%	0.05%	-4.8007	0.4902
TF binding+any motif+DNase peak	138492	44.90%	35.59%	13.16%	0.13%	-4.7748	0.4958
TF binding+matched TF motif+DNase peak	7691	47.12%	33.86%	11.73%	0.05%	-4.7943	0.4914
TF binding or DNase peak	2281669	50.99%	29.70%	10.71%	0.11%	-4.8194	0.5030
eQTL+TF binding+matched TF motif+DNase peak	70	51.43%	31.43%	11.43%	0.00%	-4.7233	0.4878
TF binding+matched TF motif	6873	58.74%	24.12%	8.29%	0.06%	-4.8432	0.5015

392 Supplementary Figures

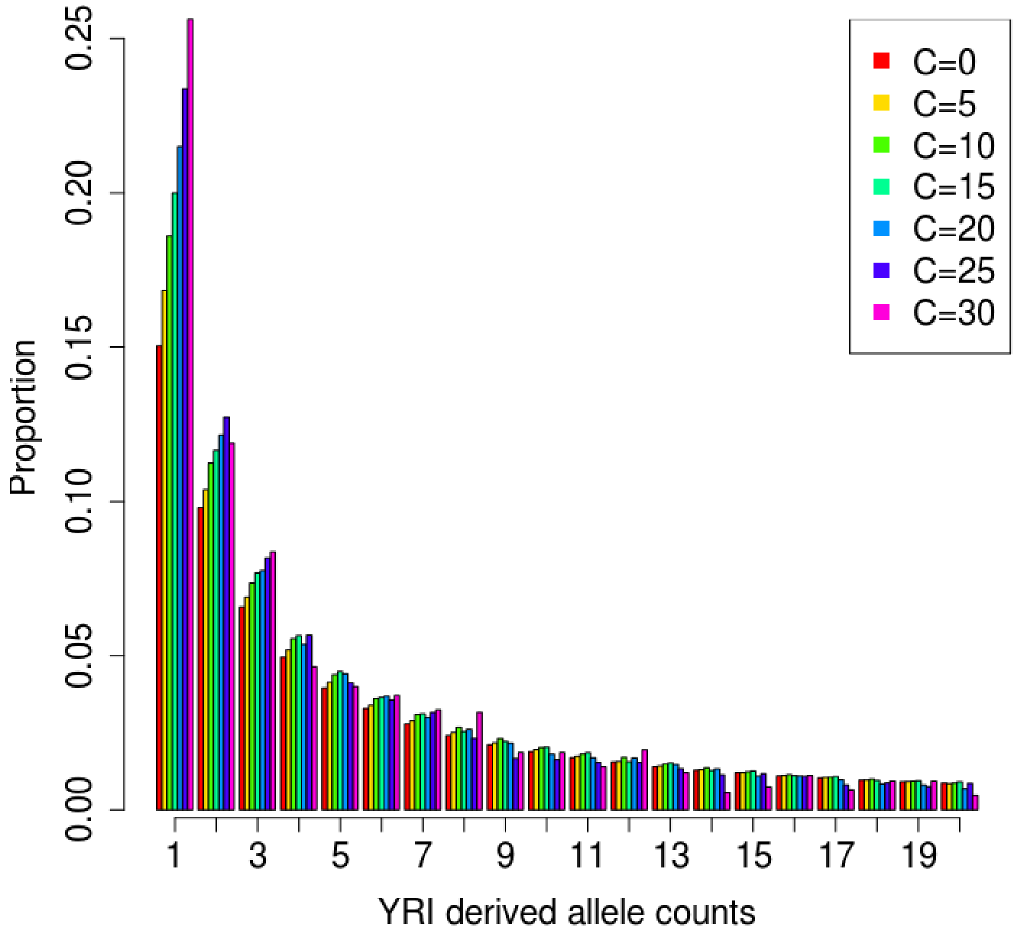

Figure S1. First 20 bins of the observed SFS for sites under different C-score bins. Note that the spectrum gets more skewed towards singletons with increasing C-scores, likely reflecting the action of negative selection on deleterious mutations.

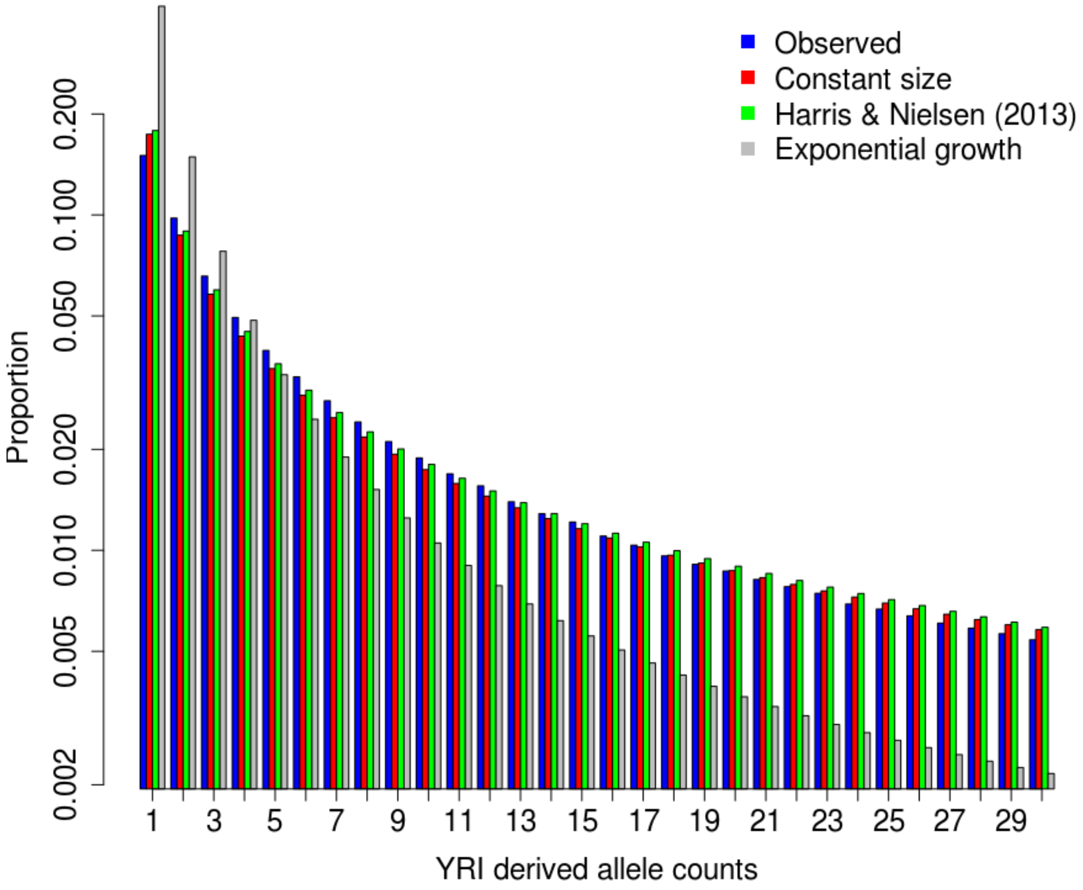

Figure S2. First 30 bins of the observed SFS for sites with C=0 (blue). The full SFS was fit to different models of neutral evolution under the Harris and Nielsen (2013) model (green), a model of constant size (red) or an exponentially growing population size model (here only shown running for t=10,000 generations at rate 5, grey). The y-axis is on a log-scale. The best-fitting exponential growth model was the one with the smallest rate (1) and duration (1,000 generations) and looked similar to the constant and Harris and Nielsen models, but was still not as good a fit as either of the latter two.

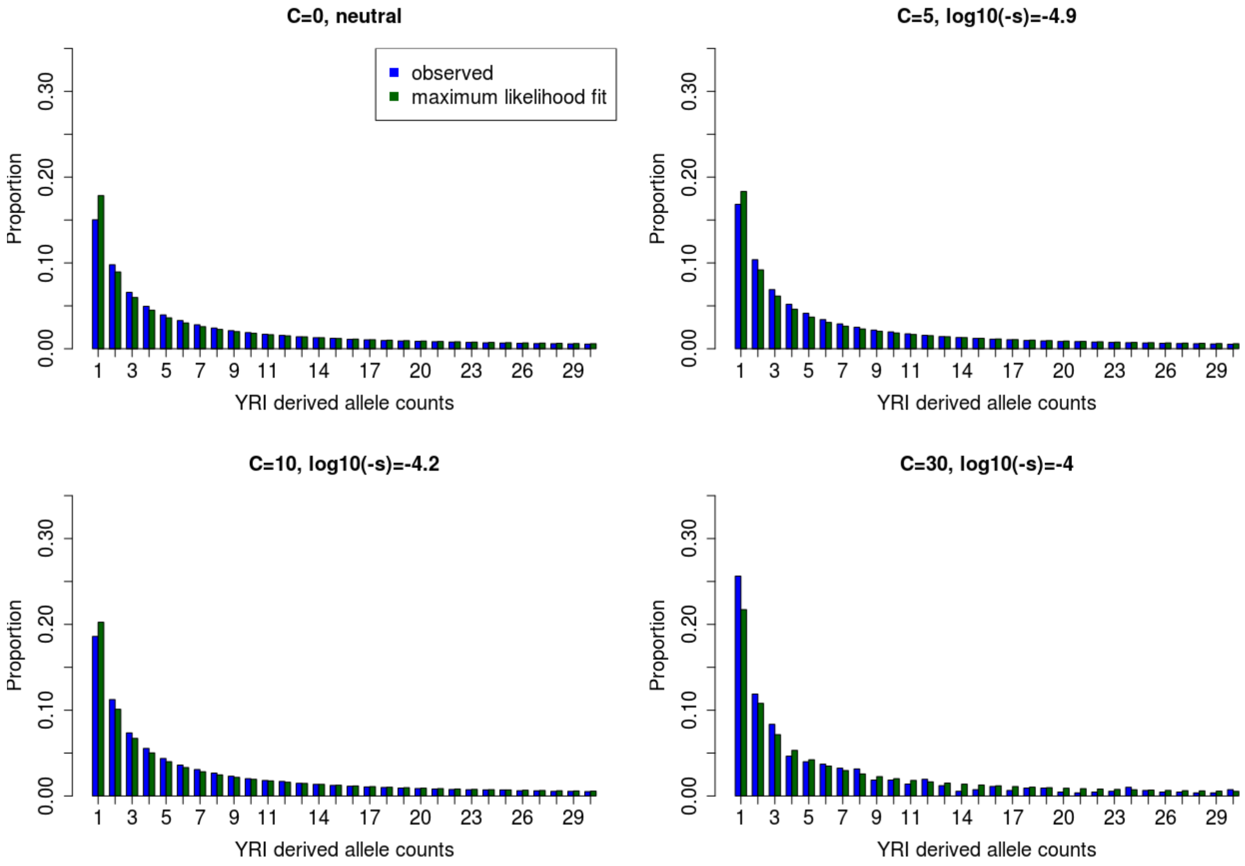

Figure S3. First 30 bins of the observed SFS for a few representative C-score bins and their corresponding maximum likelihood selection models.

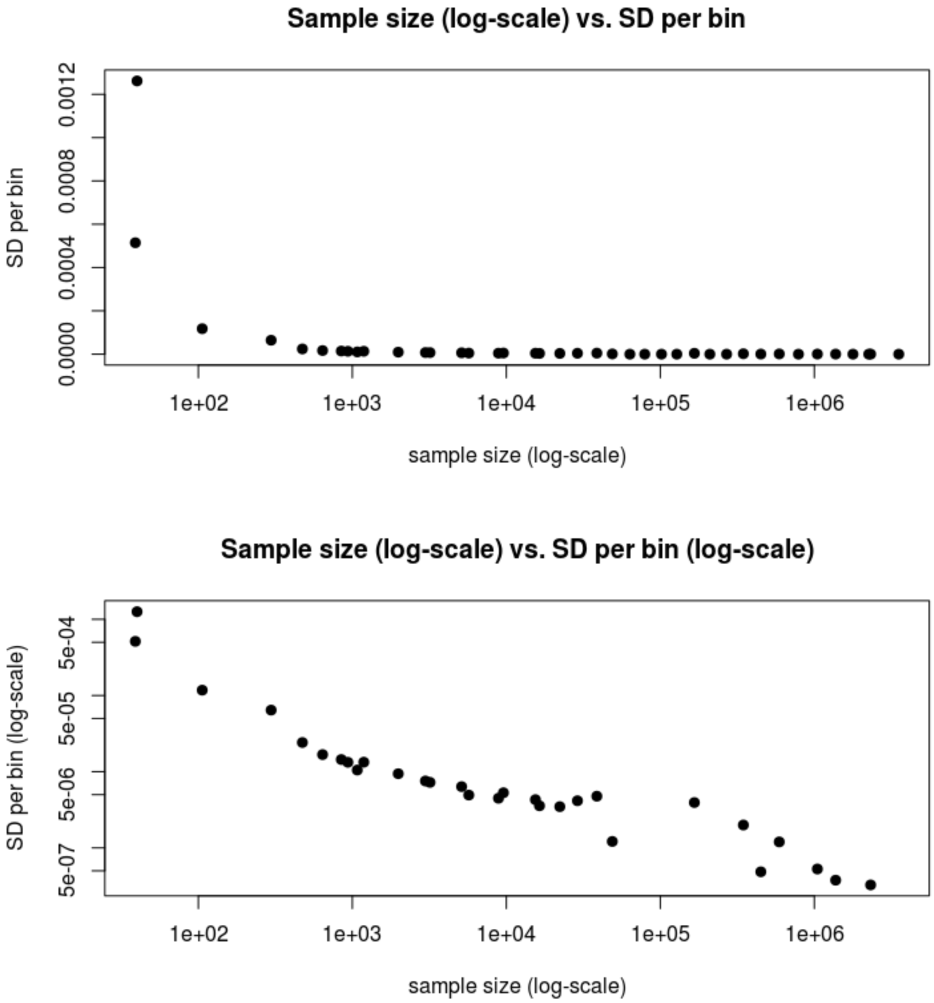

Figure S4. Comparison of standard deviations and size of bins. Top panel: Standard deviation per C-score bin plotted as a function of sample size per bin (log-scale). Bottom panel: Same plot but with the y-axis on a log-scale.

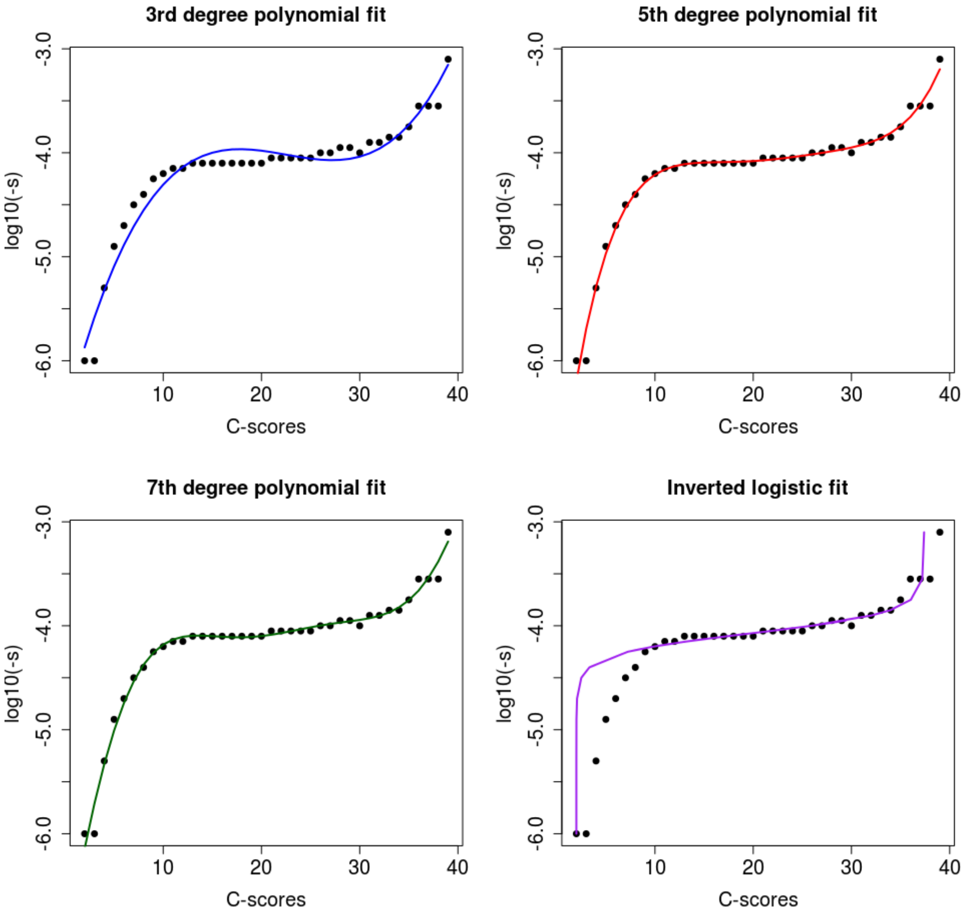

Figure S5. Fitting of different functions to C-score mappings. We attempted to fit polynomial functions to log(-s) as a function of C-scores and a logistic function to C-scores as a function of log(-s). We find that the polynomial functions are a better fit than the logistic function, and, among the polynomial functions, the 5th degree polynomial (with a sum of least squares = 0.1962) is the only one that is both monotonically increasing and not showing signs of overfitting.

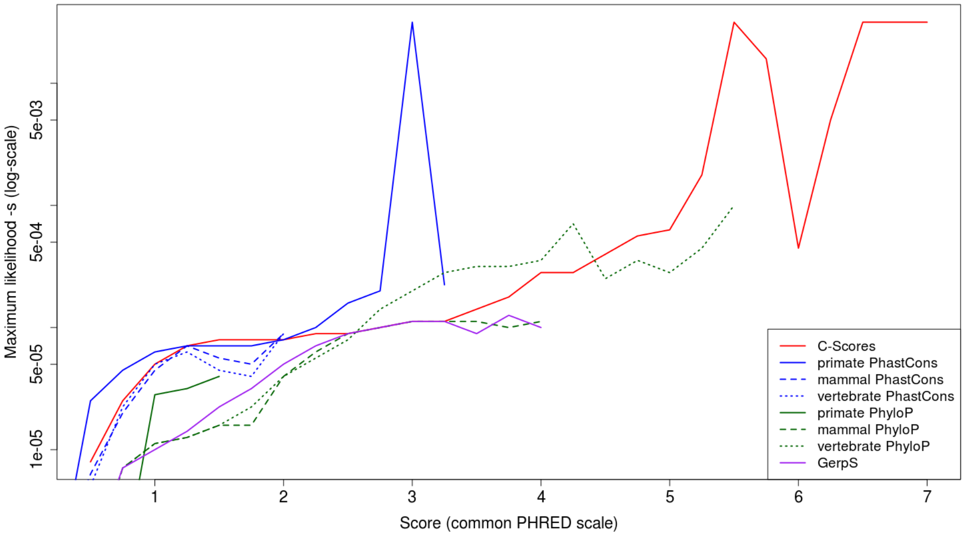

Figure S6. Maximum likelihood mapping of different types of scores to a selection coefficient scale, excluding bins mapped to neutrality. Before mapping, scores were re-scaled on a common PHRED scale (see main text). The wide fluctuations to the right of the image are due to the small number of sites per bin at highly deleterious bins. We exclude these bins when fitting C-scores to selection coefficients in our main analysis.

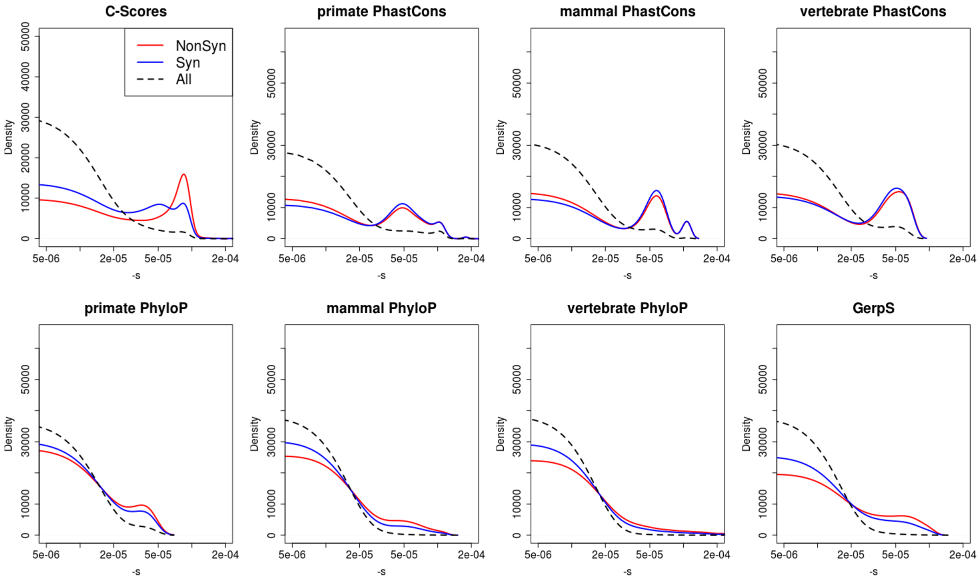

Figure S7. Distribution of fitness effects at nonsynonymous, synonymous and all polymorphisms in Yoruba, using different types of conservation scores for mapping. We note that some form of bimodality at coding sites is observed in all but one of the distributions.

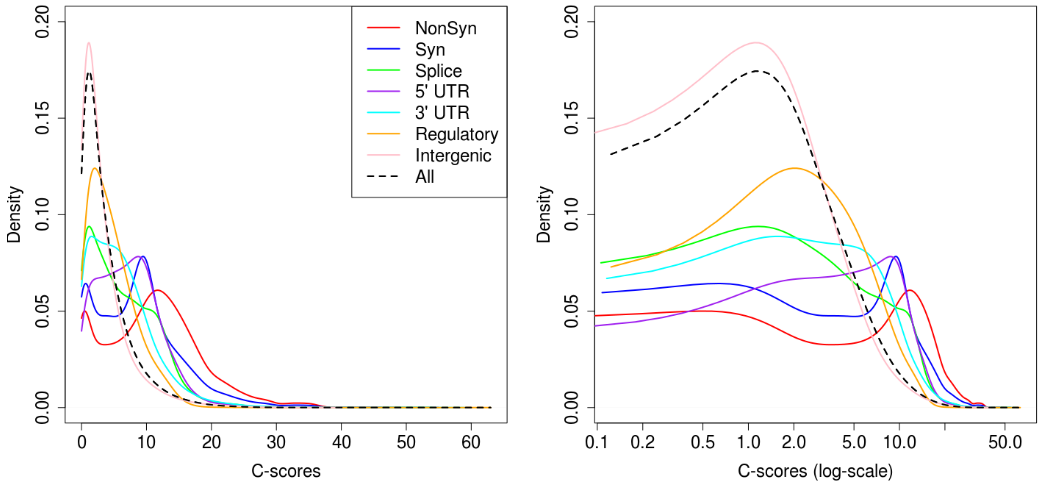

Figure S8. Distribution of unmapped C-scores among YRI polymorphisms, partitioned by the genomic consequence of the mutated site. Consequences were determined using the Ensembl Variant Effect Predictor (v.2.5). NonSyn = nonsynonymous. Syn = synonymous. Splice = splice site.

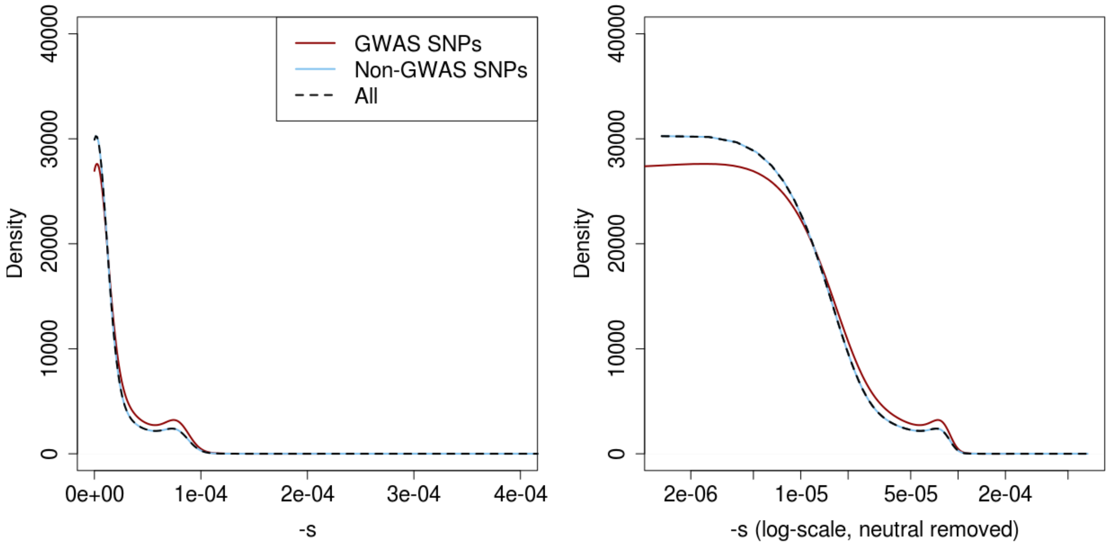

Figure S9. Distribution of fitness effects among YRI polymorphisms, partitioned by whether the SNPs are found in the GWAS database or not. The right panel shows a zoomed-in version of the same distributions after removing neutral polymorphisms and log-scaling the x-axis.

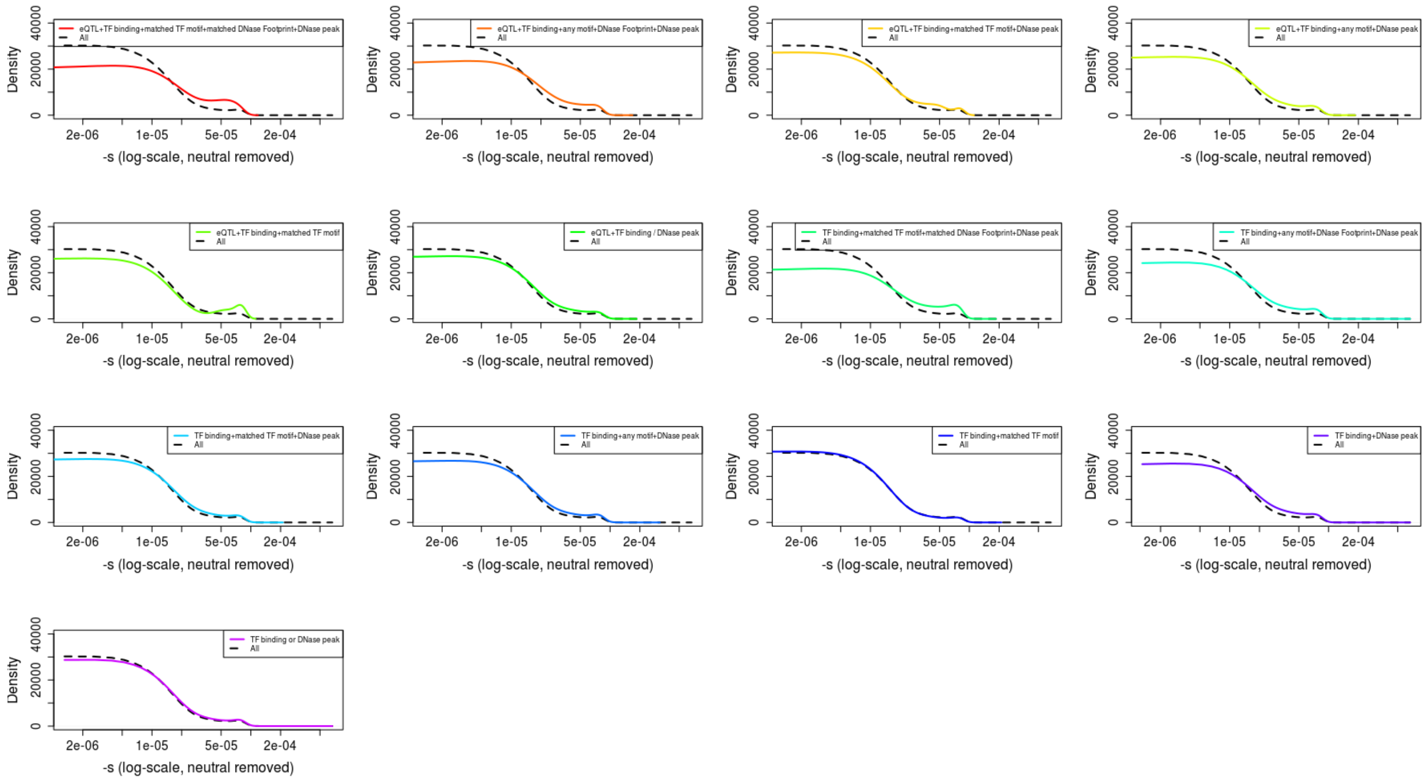

Figure S10. Distribution of fitness effects among different types of RegulomeDB regulatory YRI polymorphisms, obtained from various ENCODE assays. The black dashed line corresponds to the distribution of all YRI SNPs.