Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation

Published Version
doi:10.1371/journal.pgen.1000083

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:4728747

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome

Adam R. Boyko1,2, Scott H. Williamson1, Amit R. Indap1, Jeremiah D. Degenhardt1, Ryan D. Hernandez1, Kirk E. Lohmueller1,2, Mark D. Adams3, Steffen Schmidt4, John J. Sninsky5, Shamil R. Sunyaev4, Thomas J. White5, Rasmus Nielsen6, Andrew G. Clark2, Carlos D. Bustamante1*

1 Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, United States of America, 2 Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, United States of America, 3 Department of Genetics, BRB-624, Case Western Reserve University, Cleveland, Ohio, United States of America, 4 Division of Genetics, Department of Medicine, Brigham & Women’s Hospital and Harvard Medical School, Boston, Massachusetts, United States of America, 5 Celera Diagnostics, Alameda, California, United States of America, 6 Center for Comparative Genomics, University of Copenhagen, Copenhagen, Denmark

Abstract

Quantifying the distribution of fitness effects among newly arising mutations in the human genome is key to resolving important debates in medical and evolutionary genetics. Here, we present a method for inferring this distribution using Single Nucleotide Polymorphism (SNP) data from a population with non-stationary demographic history (such as that of modern humans). Application of our method to 47,576 coding SNPs found by direct resequencing of 11,404 protein coding-genes in 35 individuals (20 European Americans and 15 African Americans) allows us to assess the relative contribution of demographic and selective effects to patterning amino acid variation in the human genome. We find evidence of an ancient population expansion in the sample with African ancestry and a relatively recent bottleneck in the sample with European ancestry. After accounting for these demographic effects, we find strong evidence for great variability in the selective effects of new amino acid replacing mutations. In both populations, the patterns of variation are consistent with a leptokurtic distribution of selection coefficients (e.g., gamma or log-normal) peaked near neutrality. Specifically, we predict 27–29% of amino acid changing (nonsynonymous) mutations are neutral or nearly neutral (|s|<0.01%), 30–42% are moderately deleterious (0.01%<|s|<1%), and nearly all the remainder are highly deleterious or lethal (|s|>1%). Our results are consistent with 10–20% of amino acid differences between humans and chimpanzees having been fixed by positive selection with the remainder of differences being neutral or nearly neutral. Our analysis also predicts that many of the alleles identified via whole-genome association mapping may be selectively neutral or (formerly) positively selected, implying that deleterious genetic variation affecting disease phenotype may be missed by this widely used approach for mapping genes underlying complex traits.

Introduction

The distribution of fitness effects of newly arising mutations largely determines the importance of different evolutionary forces (e.g., strong and weak selection, genetic drift, recombination, and bottlenecks) on the patterning the history of a population [1]. Quantifying this distribution could, therefore, inform debates in evolutionary and medical genetics including the advantages of sex and recombination [2], the applicability of neutral and nearly neutral theories of molecular evolution to population genetics [3,4], and the importance of different population genetic theories of human complex disease [5]. Estimates for the proportion of mutations in the human genome that are neutral, deleterious, and nearly-neutral vary widely [6–11]. Many factors account for differences among estimates including quantity and quality of data, methodological approaches, and interpretation of results (including definitions of what is meant by deleterious, neutral, and nearly-neutral).

Two main approaches are used to estimate the distribution of fitness effects (DFE) of new mutations: direct estimates based on mutation accumulation (MA) experiments and indirect estimates based on population-genetic analyses of polymorphism and/or divergence data. MA experiments have generally found that the fitness distributions of new mutations are consistent with a leptokurtic (i.e., highly peaked) DFE with a maximum at neutrality [12–15]. The precise functional form (e.g., gamma, lognormal, etc.) remains unknown, however, since MA experiments are labor intensive and can only measure the fitness consequences of mutations of large effect. Another drawback is that MA cannot readily be applied to many species of interest (including humans) for both ethical and practical reasons.

Analyses of McDonald-Kreitman (MK) contingency tables (i.e., comparison of the silent to replacement ratio of polymorphisms and fixed differences) have yielded various conclusions about the DFE, including that it may be normally distributed [16], nearly exponentially distributed [17], or gamma distributed with a shape...
Author Summary

Although mutations are known to cause varying degrees of harmful effects, it is difficult to quantify the distribution that best describes the variation of fitness effects of these mutations. Here we present a new method for inferring this distribution and inferring population history using Single Nucleotide Polymorphism (SNP) data from human populations. Using 47,576 SNPs discovered in 11,404 genes from sequencing 35 individuals (20 European Americans and 15 African Americans), we find evidence of an ancient population expansion in the sample with African ancestry and a relatively recent bottleneck in the sample with European ancestry. In both populations, the patterns of variation are consistent with a leptokurtic distribution of selection coefficients (e.g., gamma or log-normal) peaked near neutrality. Specifically, we predict 27–29% of amino acid changing (nonsynonymous) mutations are neutral or nearly neutral, 30–42% are moderately deleterious, and nearly all the remainder are highly deleterious or lethal. Furthermore, we infer that 10–20% of amino acid differences between humans and chimpanzees were fixed by positive selection, with the remainder of differences being neutral or nearly neutral.

distributions over a range of possible fitness effects, weighted according to the candidate distribution, to find the maximum likelihood distribution of fitness effects. As in previous studies (e.g., [1,10,23]), we focus on the DFE of deleterious nonsynonymous mutations. We assume s and 2s are the selective disadvantage of heterozygotes and homozygous mutants and infer the DFE based on the scaled selection parameter, $\gamma = 2s$. Our inference method only considers polymorphism data; however, once we have inferred the demographic and selective parameters, we can calculate the expected number of nonsynonymous fixed differences between humans and chimpanzees. Comparing our estimate to the number of observed differences offers a means of testing the predictions of our model.

We applied this method to a database of human autosomal coding SNPs ascertained without bias and for which a syntenic, informative chimpanzee outgroup nucleotide was available. SNPs were initially obtained by direct resequencing of exons from 20,362 putative genes in 15 African Americans and 20 European Americans. In total, 17.8 Mb of genic sequence (approximately 61% of autosomal ReSeq gene bases) containing 47,576 diallelic SNPs (25,145 synonymous and 22,431 nonsynonymous) from 11,404 genes met our bioinformatic criteria and were used in our population analysis.

Results

Numerical and Simulation Results To Test Performance of the Method

The statistical method described below was implemented in an ANSI C computer program which estimates parameters of multi-epoch population size change models, and conditional on demographic parameters, estimates parameters of 13 different selection models (see Table 1). To test the accuracy of the program we simulated 100 datasets with and without linkage under the best-fit demographic model and gamma-distribution selection model inferred from the data (see Methods). The method was unbiased in estimating the demographic and selection parameters, and the amount of linkage in the dataset did not affect the accuracy or power of our inference appreciably (Figure 1).

Inference of Demographic History

We find that a two-epoch instantaneous growth model provides a good fit to the synonymous site-frequency spectrum in the African American sample ($ \chi^2_{\text{obs}} = 40.3; p > 0.05$ based on coalescent simulations with recombination, see Figure S1, Figure S2, Figure S3 and Table S1). For the European American data, in contrast, growth or single bottleneck models were a poor fit ($p < 0.001$). In order to even marginally fit the European American SFS data, a model with at least six parameters is necessary (i.e., the “complex” bottleneck scenario; $ \chi^2_{\text{obs}} = 133.5; p \sim 0.01$). The reason for this is that the European Americans show a large number of high-frequency derived SNPs even after correcting for multiple-hits using the method of Hernandez et al. [22]. This uptick in high frequency derived SNPs is also evident in the European American nonsynonymous site-frequency spectrum and because of the uncertainty in the cause of the departure from the model at synonymous sites, we focus much of our analyses on the African American data which showed no such bias.

Distribution of Fitness Effects of Newly Arising Mutations

To infer the DFE at nonsynonymous sites, we analyzed the unfolded nonsynonymous site-frequency spectra using 13 different selection models (see Table 1 and Figure 2). After demographic correction (see Methods and Text S1), both populations showed
Table 1. Maximum likelihood estimates of proposed distributions of deleterious fitness effects of new nonsynonymous mutations.

<table>
<thead>
<tr>
<th>Model</th>
<th>ALL</th>
<th># fixed</th>
<th>df distribution</th>
<th>MLE (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>8536.2</td>
<td>86,897</td>
<td>0 Pr(γ = 0) = 1</td>
<td>k = −7.324 (−7.86, −6.81)</td>
</tr>
<tr>
<td>Fixed (pt mass)</td>
<td>2801.4</td>
<td>4,581</td>
<td>1 Pr(γ = k) = 1</td>
<td>λ = 0.0365 (0.0336, 0.0400)</td>
</tr>
<tr>
<td>Exponential</td>
<td>757.1</td>
<td>7,894</td>
<td>1 Pr(γ = −x) = EXP(λ)</td>
<td>p = 0.3224 (0.314, 0.331)</td>
</tr>
<tr>
<td>Neutral-lethal</td>
<td>136.0</td>
<td>28,016</td>
<td>1 Pr(γ = 0) = p^0; Pr(γ = −x) = 1 − p^0</td>
<td>p = 0.184 (0.158, 0.206), k = −2.77</td>
</tr>
<tr>
<td>Normal</td>
<td>225.9</td>
<td>65,078</td>
<td>2 Pr(γ = x) = NORM(μ, σ)</td>
<td>μ = −38.5 (−43.5, −34.0), σ = 28.6 (25.0, 32.5)</td>
</tr>
<tr>
<td>Pt mass-lethal</td>
<td>44.0</td>
<td>17,878</td>
<td>2 Pr(γ = k) = p; Pr(γ = −x) = 1 − p</td>
<td>p = 0.372 (0.358, 0.387), k = −1.79 (−2.12, −1.45)</td>
</tr>
<tr>
<td>Exponential+lethal</td>
<td>28.6</td>
<td>17,754</td>
<td>2 Pr(γ = −x) = p*EXP(λ); Pr(γ = −x) = 1 − p</td>
<td>λ = 0.373 (0.298, 0.464), p = 0.392 (0.375, 0.415)</td>
</tr>
<tr>
<td>Exponential+neutral</td>
<td>7.0</td>
<td>22,133</td>
<td>2 Pr(γ = 0) = p^0; Pr(γ = −x) = (1 − p^0)*EXP(λ)</td>
<td>λ = 0.0048 (0.0037, 0.0061), p = 0.245 (0.231, 0.256)</td>
</tr>
<tr>
<td>Lognormal</td>
<td>5.1</td>
<td>19,812</td>
<td>2 Pr(γ = −x) = LOGNORM(μ, σ)</td>
<td>μ = 5.02 (4.45, 5.50), σ = 0.361 (5.18, 6.74)</td>
</tr>
<tr>
<td>Gamma</td>
<td>3.7</td>
<td>20,113</td>
<td>2 Pr(γ = −x) = GAMMA(μa, β)</td>
<td>μ = 0.184 (0.158, 0.206), β = 8200 (3500, 20300)</td>
</tr>
<tr>
<td>Neutral+pt mass-lethal</td>
<td>3.8</td>
<td>21,335</td>
<td>3 Pr(γ = 0) = p^0; Pr(γ = k) = p; Pr(γ = −x) = 1 − p^0 = p</td>
<td>p = 0.245 (0.222, 0.266), p = 0.208 (0.176, 0.294), k = −13.3 (−8.8, −25.3)</td>
</tr>
<tr>
<td>Neutral+gamma</td>
<td>2.9</td>
<td>20,758</td>
<td>3 Pr(γ = 0) = p^0; Pr(γ = −x) = (1 − p^0)*GAMMA(μa, β)</td>
<td>p = 0.184 (0.0, 0.235), μ = 0.344 (0.178, 0.790), β = 1900 (280, 12300)</td>
</tr>
<tr>
<td>Neutral-exponential+lethal</td>
<td>2.7</td>
<td>20,956</td>
<td>3 Pr(γ = 0) = p^0; Pr(γ = −x) = (1 − p^0)*EXP(λ); Pr(γ = −x) = p</td>
<td>λ = 0.02818 (0.0085, 0.072), p^0 = 0.2176 (0.178, 0.245), p = 0.4525 (0.200, 0.548)</td>
</tr>
<tr>
<td>Normal+lethal</td>
<td>–</td>
<td>24,300</td>
<td>3 Pr(γ = x) = p*EXP(μa, σ); Pr(γ = −x) = 1 − p</td>
<td>p = 0.428 (0.406, 0.458), μ = 4.44 (−5.40, −3.55), σ = 4.44 (4.4, 6.5)</td>
</tr>
</tbody>
</table>

ML estimates and predicted number of human-chimp fixed differences under each model computed after applying demographic correction. Distributions are in terms of the scaled selection coefficient, γ = 2Ncurrs, where Ncurrs is 25,636 in African Americans and 52,907 in European Americans. ALL is the likelihood difference between the model and the overall best-fit model for the population; # fixed is the number of nonsynonymous fixed differences predicted by the model. Approximate 95% confidence intervals based on semi-parametric bootstrap are reported for African American parameter estimates.

doi:10.1371/journal.pgen.1000083.t001

less diversity than expected under strict neutrality at nonsynonymous sites, especially at high-frequency derived classes, indicative of purifying selection (Figure 2).

One-parameter models of selection, particularly those assigning a single selection coefficient to all nonsynonymous mutations, did a poor job of recovering the nonsynonymous site-frequency spectrum and predicting the number of observed nonsynonymous differences (Table 1). The best one-parameter model was a model of purifying selection which predicts 32.2% of amino acid sites are neutral, and 68.8% are strongly constrained. This model is 136 log-likelihood units below the maximum of more complex two- and three-parameter models, suggesting additional parameters provide highly significant improvements to the models.
Several two-parameter selection models—notably the gamma and lognormal—resulted in much better predictions ($D^2_{LL} = 3.7$ and $D^2_{LL} = 5.1$ below the maximum). All of these best-fit two-parameter selection models slightly under-predicted the number of observed nonsynonymous human-chimp differences which is consistent with the fixation of rare positively-selected mutations that were not incorporated into these models. Three-parameter selection models did not provide a measurably improved fit to the site-frequency spectrum except for the normal distribution model with a proportion of lethal site ($normal + lethal$, $D^2_{LL} = 0$), which was the only model tested that allowed for weak positive selection on segregating variation.

The best fitting gamma and lognormal distributions are both highly leptokurtic, with the bulk of mass centered on neutrality and a long negative tail extending into lethality. Our confidence in the proportion of nearly neutral mutations ($\mid s \mid < 0.0001$) is tightly clustered around 28% in African Americans (27.3–29.0%, Tables 2 and 3). Our inferences are less tightly centered for more deleterious classes, as expected since the estimates of these classes are basically extrapolations of the distribution from the nearly-neutral data, since highly deleterious classes contribute little if any to observed polymorphism. Our confidence intervals incorporate both our demographic and selection parameter uncertainty, although they do critically depend on the selective effects following the gamma distribution. For example, the normal+lethal model also provides a good fit to the data but predicts somewhat different proportions of mutations in each selective class (Figure 3). This illustrates a general point that methods that make use of standing genetic variation to infer the strength of selection have little power to distinguish among models for the strength of highly deleterious mutations since these mutations contribute little to extant polymorphism. We note that in African Americans, the mean selection coefficient for newly arising mutations under the best-fit distributions is $-0.058$ for the lognormal and $-0.029$ for the gamma (95% C.I. = $[-0.059, -0.018]$), but the mean selection coefficient for mutations that are segregating (have two alleles in the sample) is less negative and nearly identical for the two distributions ($-0.000136$ for the lognormal and $-0.000140$ with 95% C.I. = $[-0.000143, -0.000134]$ for the gamma). Recently, Kryukov et al. [25] have estimated $s$ for rare segregating human alleles, and their estimated range ($-0.003 -- -0.001$) is, encouragingly, between our estimate for new mutations and for segregating alleles.

Although selective inference is complicated by the complex demographic history of European Americans, we find no evidence that the DFE among new mutations differs between European and African Americans if we assume a gamma or lognormal distribution of fitness effects (Table 2; Figure S4). Other
distributions provide a better fit to the European American data, but these all have a neutral point mass or a positive selection component which may lead to overfitting to the uptick of high-frequency derived SNPs observed in this population. We find it highly unlikely that the true DFE lacks moderately deleterious mutations, as is predicted by the best-fitting exponential+neutral or normal+gamma models inferred from the European American data (Table 2). We infer that European Americans have a similar mean selection coefficient for newly arising mutations as African Americans (0.030 under the best-fit gamma model), but that the

Figure 2. Observed and expected nonsynonymous site-frequency spectra after demographic correction. Expected site-frequency spectra under best-fit selection models after demographic correction. Note the logarithmic scale of the y-axis. (A) African American replacement SNPs versus expectation under neutrality, fixed selective effects, and gamma distribution of fitness effects. (B) European American replacement SNPs versus expectation under neutrality, fixed selective effects, and gamma distribution of fitness effects. doi:10.1371/journal.pgen.1000083.g002

Table 2. Distribution of fitness effects under various best-fit selective models.

<table>
<thead>
<tr>
<th>Population ancestry</th>
<th>African</th>
<th>African</th>
<th>European</th>
<th>European</th>
<th>European</th>
<th>African</th>
<th>European</th>
<th>European</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gamma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lognormal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00001&lt;</td>
<td>s</td>
<td>&lt;0.0001</td>
<td>27.9%</td>
<td>28.4%</td>
<td>24.3%</td>
<td>24.6%</td>
<td>25.0%</td>
</tr>
<tr>
<td></td>
<td>0.0001&lt;</td>
<td>s</td>
<td>&lt;0.01</td>
<td>14.7%</td>
<td>14.3%</td>
<td>14.7%</td>
<td>15.7%</td>
<td>15.7%</td>
</tr>
<tr>
<td></td>
<td>0.01&lt;</td>
<td>s</td>
<td>&lt;0.1</td>
<td>21.9%</td>
<td>15.4%</td>
<td>23.1%</td>
<td>17.5%</td>
<td>53.7%</td>
</tr>
<tr>
<td></td>
<td>s</td>
<td>&gt;0.1</td>
<td>35.9%</td>
<td>41.9%</td>
<td>37.9%</td>
<td>42.3%</td>
<td>20.5%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Proportion of mutations falling into each selection interval for each population under each best-fit models. Demographic model parameters are listed in Table 3; selection model parameters are listed in Table 1 except for African American stationary gamma (τ = 0.55, β = 37), and European American stationary gamma (τ = 0.36, β = 120), and European American simple bottleneck gamma (τ = 0.228, β = 5200).

doi:10.1371/journal.pgen.1000083.t002
mean selection coefficient for segregating mutations in the sample is more negative in European Americans (−0.000344) than in African Americans (−0.000140). It is important to note that these estimates are likely associated with wide confidence intervals due to the complex bottleneck history in European-Americans (the large number of demographic parameters in the European-American model makes it computational infeasible to calculate the confidence intervals as we did for African-Americans).

Our estimates of the proportion of mutations that are strongly deleterious ($|s| > 0.0001$), mildly deleterious (0.0001 < $|s|$ < 0.001), weakly deleterious (0.001 < $|s|$ < 0.01), and nearly neutral ($|s|$ < 0.01) are not affected by using the folded (i.e., the distribution of minor allele frequency) vs. unfolded site-frequency spectrum or by removing singletons from the analysis (Table 3). This suggests that biases in SNP calling (which would predominantly affect singleton SNPs) and polarization of the ancestral state of mutations (which would only affect the unfolded SFS) have little influence on our inferences from the African American data. Furthermore, reasonable levels of weak, purifying selection at putatively neutral synonymous sites have only a modest affect on the inference of selection at nonsynonymous sites, although they can alter the demographic parameters significantly (Table 3).

### Distribution of Fitness Effects on Nonsynonymous SNPs and Fixed Differences

Based on our inference of the DFE of newly arising mutations, we can estimate the proportion of segregating SNPs or human-chimp fixed differences that fall within various selective ranges. Assuming the best-fit two-parameter gamma-distributed selection model, we estimate that around half of nonsynonymous mutations are strongly or mildly deleterious ($|s| > 0.01$). Because of the strength of purifying selection, however, we estimate that fewer than 0.5% of amino acid replacing SNPs segregating at any frequency above 5% have fitness effects this extreme (see orange
bars in Figure 4). That is, most of the segregating variation above 5% frequency in the population is predicted to be nearly neutral (|s|<0.01%; Figure 4 green and blue bars) or weakly deleterious (0.01%<|s|<0.1%; Figure 4 yellow bars) with a higher proportion of neutral variation as the allele frequency increases. This culminates in a prediction that only 1% of nonsynonymous human-chimp differences are slightly deleterious (yellow bars), with the remainder being neutral or positively selected (Figure 4).

Our data do show significant evidence for positive selection having played an important role in patterning human-chimp amino acid divergence. All best-fit two- and three-parameter selection models which do not include positive selection (i.e., all except Normal+Hethal) underpredict the number of human-chimp nonsynonymous differences, consistent with a portion of differences being fixed through positive selection. For example, the expected number of nonsynonymous differences under the African American gamma model (20,113) is 9.3% less than what we observe (22,180; 95% C.L. = 5.4%, 12.8%). Modeling the DFE as a gamma function plus a positive selection component (Figure 4A–C) results in a significantly better fit to the nonsynonymous site-frequency spectrum than the best-fit gamma function (ΔLL = 1; ΔLL = 4.1 for model in Figure 4B) or the best-fit gamma function with a neutral point mass (ΔLL = 0; ΔLL = 3.3 for model Figure 4B).

Furthermore, including positive selection of varying strength (from weak to moderate to strong) into the model explicitly (Figure 4A–C) increases our estimate of the proportion of differences fixed through adaptive evolution to around 20% instead of 9.3% but does not alter our deleterious mutation estimates appreciably. Based on these inferences, we expect that 10–20% of human-chimp differences were fixed through positive selection, depending on the exact nature of the DFE of beneficial mutations and the relative proportion of weakly versus strongly beneficial mutations. It is also important to note that if the strength of selection of beneficial mutations is very small (e.g., γ+<5), then our approach predicts that more than 20% of human-chimp differences are adaptive and that appreciable levels of human segregating variation are subject to positive selection (Figure 4A).

**Concordance of PolyPhen Classifications and Inference of Selection**

To further refine our estimates, we utilized PolyPhen [26,27] to classify amino acid replacing SNPs as “benign”, “possibly damaging” or “probably damaging” based on site-specific sequence conservation among mammals as well as location in the three dimensional structure of the protein molecule (if known). (It is important to note that the term “damaging” used by PolyPhen is meant to reflect only that the mutation affects protein structure and not that the mutation results in loss or gain of function). After running PolyPhen, our dataset was reclassified into 15,916 benign, 4,199 possibly damaging, and 2,646 probably damaging SNPs, and we inferred the DFE for each class. As
expected, the degree of purifying selection increases from benign to possibly damaging to probably damaging (Table 4). For possibly and probably damaging SNPs, the lognormal distribution was the best-fitting two-parameter model; for benign SNPs, the neutral-exponential distribution provided the best fit (Table 4). We estimate that approximately 5% of nonsynonymous differences classified by PolyPhen as benign were fixed through positive selection, while 27% and 35% of those classified as possibly and probably damaging were inferred to be fixed through positive selection. Furthermore, although the selection coefficient of a segregating benign mutation is less negative than that of a segregating damaging mutation, because there are so many more benign mutations and because they segregate at higher frequencies, the cumulative effect of benign mutations on fitness is greater than the cumulative effect of the rarer but more damaging mutations. Based on the best-fit models in Table 4 and the observed allele frequencies in each class, we predict the average fitness in the population is reduced by 13.4% due to weakly deleterious “benign” alleles but only by 4.3% and 3.2% for possibly and probably damaging alleles, respectively.

### Discussion

This study represents the most comprehensive analysis to date using polymorphism frequencies to infer the DFE of newly arising mutations in the human genome. Several previous studies have suggested that the DFE is highly leptokurtic with a sharp peak around neutrality and a long tail extending into lethality. However, we find that such models provide a good fit to the data, but so does a mixture model with a Gaussian component centered on moderately deleterious mutations (~45%) and 55% of mutations being lethal.

Our estimate that 27.3–29.0% of nonsynonymous mutations are neutral is remarkably consistent (although somewhat larger) than Eyre-Walker et al.’s [11] estimate of 23% [21–29%] inferred from a smaller dataset using folded frequency data and the hybrid approach of Yampolsky et al. [10] except that we expect a slightly larger proportion of mutations to be nearly neutral (Table 5). Using the folded site-frequency spectrum (i.e., minor allele frequency distribution) or removing singleton SNPs had little effect on the predictions of the model (Table 4), suggesting that they are robust to errors in SNP calling, assembly, and alignment to the chimpanzee genome. Even if selection at synonymous sites were as strong as $\gamma = -1$ (stronger than current estimates in humans; e.g., [28,29]), the parameter estimates for the fitness distribution are within those estimated under the neutral assumption (Table 3). The demographic inference, however, changes if there is selection at silent sites—we infer a significantly larger population size undergoing an expansion event significantly more recently than we did under the assumption of synonymous sites being neutral (Table 3). Since $\gamma = -1$ is almost certainly an overestimate of selection on synonymous sites, and since the resulting model results in a poorer fit to both the synonymous and nonsynonymous site-frequency spectra ($\Lambda L_{Sil} = 14.4$ and $\Lambda L_{Rep} = 14.4$ in African Americans), our inferences using $\gamma = 0$ at synonymous sites should be reasonable.

It is important to note that our approach differs from some recent methods proposed for estimating the DFE on new mutations as well as some recent estimates of the DFE in humans (e.g., [24]). That algorithm, as described and implemented, only considers a single size change event due to computational limitations of their matrix multiplication approach. Our method, on the other hand, uses an algorithm with running time independent of the number of size change events, so that is effectively no limit to the number of events that can be considered (e.g., the population size can change every generation with essentially no added overhead). Our conclusions also differ from those of the previous group in that we find no significant difference in the DFE between African and European Americans, although we do find that alleles segregating in European Americans are on average more deleterious than those segregating in African Americans consistent with Lohmueller et al. [30]. Our comparison of observed versus predicted site-frequency spectra for silent sites suggests that our three-size-change model is a significantly better fit than the single-size-change model, and that the difference in DFEs between Europeans and African Americans reported by [24] may largely be a consequence of a poor fit of the single size change model to the European SFS data. That is, demographic

### Table 4. Distribution of fitness effects in African Americans by PolyPhen class.

<table>
<thead>
<tr>
<th></th>
<th>benign</th>
<th>possibly damaging</th>
<th>probably damaging</th>
<th>combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>proportion of mutations</td>
<td>50.8%</td>
<td>25.7%</td>
<td>23.5%</td>
<td></td>
</tr>
<tr>
<td>proportion of SNPs</td>
<td>72.7%</td>
<td>17.4%</td>
<td>9.9%</td>
<td></td>
</tr>
<tr>
<td>proportion of fixed differences</td>
<td>82.4%</td>
<td>11.6%</td>
<td>6.0%</td>
<td></td>
</tr>
<tr>
<td>best-fit selection model</td>
<td>neutral-exponential</td>
<td>lognormal</td>
<td>lognormal</td>
<td>gamma</td>
</tr>
<tr>
<td>ML parameters</td>
<td>$\lambda = 0.0038$, $\mu^2 = 0.409$</td>
<td>$\mu = 4.69$, $\sigma = 2.83$</td>
<td>$\mu = 5.78$, $\sigma = 2.87$</td>
<td>$\alpha = 0.184$, $\beta = 8200$</td>
</tr>
<tr>
<td>expected num of fixed differences</td>
<td>17,954</td>
<td>1,846</td>
<td>855</td>
<td>20,655</td>
</tr>
<tr>
<td>observed num of fixed differences</td>
<td>18,272</td>
<td>2,568</td>
<td>1,341</td>
<td>22,181</td>
</tr>
<tr>
<td>$s &lt; 0.001$</td>
<td>42.0%</td>
<td>14.0%</td>
<td>7.4%</td>
<td>26.7%</td>
</tr>
<tr>
<td>$0.001 &lt; s &lt; 0.001$</td>
<td>9.3%</td>
<td>25.5%</td>
<td>18.6%</td>
<td>15.7%</td>
</tr>
<tr>
<td>$0.01 &lt; s &lt; 0.01$</td>
<td>40.2%</td>
<td>31.3%</td>
<td>30.3%</td>
<td>35.6%</td>
</tr>
<tr>
<td>$s &gt; 0.01$</td>
<td>8.4%</td>
<td>29.2%</td>
<td>43.6%</td>
<td>22.0%</td>
</tr>
<tr>
<td>mean $s$ (mutations)</td>
<td>$-0.0030$</td>
<td>$-0.0087$</td>
<td>$-0.0265$</td>
<td>$-0.0294$</td>
</tr>
<tr>
<td>mean $s$ (polymorphisms)</td>
<td>$-0.00013$</td>
<td>$-0.00027$</td>
<td>$-0.00040$</td>
<td>$-0.00014$</td>
</tr>
</tbody>
</table>

Best-fit selection distribution for each PolyPhen class of mutations in African Americans and the proportion of new mutations in each PolyPhen class falling into each selection interval. Combined values are the sum of the best-fit values across classes. All selection inferences were run under the maximum likelihood African American expansion model.

doi:10.1371/journal.pgen.1000083.t004
matters when estimating the parameters of the DFE as our and their analysis suggests and if the demographic model is misspecified, estimates of the parameters in the DFE will be significantly biased.

Despite the large number of deleterious mutations that enter the population each generation, we estimate that the vast majority of common human genetic variation (e.g., SNPs at >5% derived allele frequency) is neutral or nearly neutral. We hold that these results have important implications for interpreting the outcome of whole genome association mapping studies that aim to identify alleles underlying common human diseases. Specifically, if chronic disease has a negative impact on Darwinian fitness, whole-genome association mapping approaches that survey common genetic variation for association with disease may be missing the most evolutionarily interesting and medically relevant alleles. The reason is that these assays have higher power to identify susceptibility alleles at moderate frequency, and are therefore biased towards finding evolutionarily neutral mutations (or formerly positively selected alleles). Re-sequencing in large samples of phenotypically extreme individuals, on the other hand, is much more likely to discover rare, large-effect mutations that are predicted by our analysis (and others) to be deleterious. Population genetic theory also predicts that deleterious alleles will not spread widely from their geographic point of origin unless they stochastically drift in frequency due to founder effects or bottlenecks. Therefore, one would not expect, a priori, an association between a deleterious mutation and a disease phenotype to replicate across human populations even if the mutation has a significant attributable risk within a single population. Conversely, neutral polymorphisms (and formerly positively selected polymorphisms) that affect disease may likely replicate across human populations. These ascertainment biases in the evolutionary classes of alleles that can currently be associated with common human disease cloud our understanding of the relationship between chronic disease and Darwinian fitness.

Future work could extend this approach to allow for non-additive genetic effects and inference of more complex demographic models from joint site-frequency spectra. The method of Williamson et al. [31] of inferring selection and dominance could be incorporated into this model, although it is likely that the degree of dominance of new mutations follows a distribution whose shape may depend on the fitness effect of the mutation. Based on our results, it is unlikely that the synonymous and nonsynonymous site-frequency spectra contain enough information to allow for an inference of dominance parameters in addition to selective and demographic parameters. Like previous studies of the DFE in humans, our results should be interpreted as the distribution of heterozygote fitnesses over all genetic backgrounds [11].

A potential concern with our analysis is that we ignore linkage among selected alleles and among selected and neutral alleles in deriving predictions for the site-frequency spectrum. Ignoring linkage may bias our inference regarding the parameters of the DFE. It is important to note, however, the level of linkage in our dataset (15,681 nonsynonymous SNPs in African Americans spread over 11,404 genes and ~100,000 exons) is slight and simulation results suggest that our parameter estimates and confidence intervals are robust to this complication (Figure 1).

A consequence of this DFE is that the molecular clock should be fairly insensitive to changes in \( N_e \) at nonsynonymous sites. Assuming mutations occur at a constant rate per generation, nearly neutral theory predicts that a selective regime that is gamma-distributed with a shape parameter \( \alpha \) results in a rate of amino acid substitution proportional to \( N_e^{-\alpha} \) per generation [4,32]. Since generation time is approximately proportional to \( N_e^{0.33} \) in mammals [32], we expect amino acid replacing substitutions to fix at a rate proportional to \( N_e^{0.15} \) per year if \( \alpha = 0.158 - 0.206 \) as we have inferred. This would explain why the observed relationship between population size and substitution rate is stronger at synonymous sites (which should fix at a rate proportional to \( N_e^{0.33} \)) than at nonsynonymous sites [3,33]. However, our estimate of \( \alpha \) is sensitive to whether or not there is a point mass at neutrality in addition to a gamma distribution of fitness effects (Table 1).

Encouragingly, our frequency-based method of selective inference concurs with PolyPhen’s fitness classification method which is based on conservation and physiochemical properties and does not utilize frequency information. Benign mutations are best described by a model with a significant neutral class (40.9%) and an exponentially-distributed deleterious class, whereas possibly and probably damaging mutations are best described by lognormal distribution with significant levels of deleterious mutation (Table 4), although these last two distributions were not significantly better fitting than the gamma distribution. Despite the heterogeneity of these distributions, the overall estimate of adaptive evolution is the same as that obtained from using a single gamma distribution—both models lead to a ~10% underprediction in the number of human-chimp nonsynonymous differences. The proportionately higher levels of inferred adaptive evolution at possibly and probably damaging sites versus benign sites agrees with Kimura’s [6] observation that stochastic loss of weak beneficial mutations results in mutations of intermediate effect being most likely to adaptively fix rather than those of small effect as predicted by Fisher [34].

### Table 5. A comparison of several recent estimates of the DFE at nonsynonymous sites in humans.

<table>
<thead>
<tr>
<th></th>
<th>this study</th>
<th>Eyre-Walker et al. [11]</th>
<th>Yampolsky et al. [10]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>African American (gamma distribution)</td>
<td>European American (gamma distribution)</td>
<td>basic gamma distribution</td>
</tr>
<tr>
<td>( s &lt; 0.00001 )</td>
<td>18%</td>
<td>15%</td>
<td>11%</td>
</tr>
<tr>
<td>( 0.00001 &lt; s &lt; 0.0001 )</td>
<td>10%</td>
<td>9%</td>
<td>8%</td>
</tr>
<tr>
<td>( 0.001 &lt; s &lt; 0.01 )</td>
<td>37%</td>
<td>38%</td>
<td>37%</td>
</tr>
<tr>
<td>( s &gt; 0.01 )</td>
<td>36%</td>
<td>38%</td>
<td>44%</td>
</tr>
</tbody>
</table>

Note that the selection intervals are those used by [10] and differ from those used in Tables 2–4.

doi:10.1371/journal.pgen.1000083.t005
In this investigation we did not look directly at the distribution of positively selected amino acid mutations owing to the difficulty of inferring the distribution when there are so many fewer positively selected mutations segregating than neutral or negatively selected ones. Nevertheless, we note that both African and European Americans exhibit about 10% more amino acid divergence than expected from best-fitting selective models that ignore positive selection. This suggests 10–20% of amino acid divergence is due to positive selection (Figure 4). This result is somewhat consistent with Fay et al.’s [8] estimate that 35% of human-chimp nonsynonymous fixed differences were adaptively driven, but contrast with recent calculations suggesting essentially none of these differences were driven by positive selection [35,36]. We note that non-selective factors influencing fixation rates (such as linkage) should apply equally to synonymous sites and should not bias the number of nonsynonymous differences relative to the number of synonymous differences (divergence time was estimated from synonymous differences).

Despite the considerable dataset used in this study, our power to infer selective distributions was limited to two- and three-parameter models, some of which were indistinguishable. Although there are good theoretical and experimental reasons to prefer leptokurtic distributions like the gamma and the lognormal to fit the DFE (e.g., [4,37,38]), even with a large, genome-wide resequencing dataset, site-frequency spectrum data alone are insufficient to rule out alternative models such as the normal—lethal (Gaussian shift) model. Previous studies have noted the lack of power to infer the tails of the fitness distribution from population genetic analyses, but we see that even distributions that differ significantly at intermediate frequencies can predict nearly identical site-frequency spectra (Figure 3). Furthermore, complex demography such as that found in our European American sample (see Text S1) can limit the power to infer selection even more. Lastly, it is important to note that we have not explicitly included migration in this analysis due to computational complexity. We hope to do so in future work, and while we suspect it will improve demographic inference, we do not expect it will greatly change our estimates of the DFE.

Although we encourage adding more genes to the dataset, such work is unlikely to increase the power of our analysis substantially. Deeper resequencing within humans, on the other hand, may likely yield improvements as estimates of rare allele frequencies are refined. Likewise, if the shape of the DFE is similar across species, analyzing the DFEs in species with various population sizes could readily be solved to correct for ancestral misidentification. This phenomenon that biases selection studies using nucleotide substitution matrices for multiple-hit correction [46].

Considering each population’s corrected synonymous site-frequency spectrum separately, we used maximum likelihood to determine the best-fit demographic model. In both populations, a two-epoch model (two free parameters: timing and magnitude of size change) was a significantly better fit than a stationary model (Table S1). This two-epoch model was sufficient to account for the observed African American but not the observed European
American synonymous site-frequency spectrum (Figure S1 and Figure S2). We optimized the parameters of the African American demographic model using a multinomial likelihood function conditioning on the number of SNPs observed in the sample (i.e., each multinomial count is the proportion of SNPs in that frequency class multiplied by the total number of SNPs observed in the data), and estimated the scaled human-chimp divergence time by finding the value that correctly predicted the number of observed synonymous human-chimp differences under this demographic model assuming a stationary model of chimp demography (with $N_e$=human ancestral, $N_c$) and 300,000 generations since human-chimp divergence. We then set the European American ancestral population size and per-nucleotide mutation rate ($\mu$) to that inferred from the African American data and fit a simple bottleneck (four free parameters: timing and magnitude of bottleneck; timing and magnitude of expansion) and a complex bottleneck (six free parameters: timing and magnitude of bottleneck; timing and magnitude of recovery; timing and magnitude of expansion) using a Poisson likelihood function and back-calculated European American $N_{me}$ and $\mu$ using these parameters and a multinomial likelihood function. Although the maximum likelihood was nearly identical between the two bottleneck models and subsequent inferences of selection were robust to which of these two models we used (Table 2), the complex bottleneck provided a better goodness-of-fit, although it still failed to explain the excess of high-frequency derived SNPs in the sample. This excess may be caused by migration among Europeans and other populations or linkage to selected sites, although we cannot rule out uncorrected multiple hits or sequencing error. This excess is also apparent at European American nonsynonymous sites and is likely responsible for the overestimate in neutral and positively selected classes in some European American three-parameter selection distributions (Table 2).

**Inference of Distribution of Fitness Effects of Newly Arising Mutations**

We use the distribution of sample frequencies among variable nucleotides in the coding gene alignments (i.e., the site-frequency spectrum of coding SNPs) to infer the distribution of fitness effects among newly arising mutations. Our analytical approach makes use of standard Wright-Fisher population genetic theory within a Poisson Random Field setting [16,19,31,47–49]. The assumptions of this model include independence among SNPs, genic selection, an underlying Poisson process governing mutations, and a piecewise constant population of large size amenable to modeling using diffusion approximations. The model we employ is an extension of Williamson et al. [19], where we present the relevant population and statistical inference theory for modeling genic selection in a population experiencing a recent size change. The key addition to our previous model is a distribution of fitness effects among newly arising mutations. This amounts to modeling the components of the site-frequency spectrum—defined as $(X_1, X_2, \ldots, X_n)$ where $X_i$ is the number of SNPs with $i$ derived alleles in a population of $n/2$ individuals ($n$ chromosomes)—as independent Poisson random variables with mean:

$$E(X_i) = \frac{\theta}{2} \int_{-\infty}^{\infty} \int_{0}^{1} \cdots \int_{0}^{1} x^i (1-x)^{n-i} f(x; \Theta, \gamma) g(\gamma) dx_1 \cdots dx_n$$

where $\theta$ is the genome-wide mutation rate, $x$ represents the (unknown) population frequencies of mutations, $f(x; \Theta, \gamma)$ is the distribution of mutation frequencies given selection ($\gamma = 2N_e \lambda$) and demographic history ($\Theta$), and $g(\gamma)$ is the distribution of fitness effects among new mutations. The synonymous mutation rate ($\theta_S$) was calculated from the number of synonymous segregating sites by the method of Williamson et al. [19]; the nonsynonymous mutation rate ($\theta_K$) was fixed as $2.5 \times \theta_S$ [43]. Given the potentially different demographic histories of individuals with recent African and European ancestries, we model the two groups separately.

Although in theory $\theta = 2\lambda N_e$ can encompass any value from $\infty$ to $0$, we limited our integration of the distribution to $N_{me}$ (corresponding to lethality) to $10^{-9}$ (corresponding to neutrality), plus a point-mass (usually at $\gamma = 0$) when appropriate. Only models incorporating a normal distribution were integrated above $\gamma = 0$. For distributions with a significant (>0.01%) mass between $\gamma = -10^{-6}$ and 0, we added that mass to the weighting at $\gamma = 0$. Piecewise integrations were performed using midpoint, midinf, and qromo [30] with EPS $= 2.5 \times 10^{-9}$. Integration usually required evaluating the site-frequency spectrum at approximately 1,000 $\gamma$ values. Each evaluation took around 1 sec and the result could be stored to facilitate the evaluation of subsequent distributions. Our algorithm, implemented in C, is available in the computer program prfreq available for download from the Bustamante lab website (<http://bustamante.cb.bscb.cornell.edu/software.shml>).</p>

To obtain approximate confidence intervals around the maximum likelihood estimates for each African American selective model, we generated 200 datasets from the observed African American synonymous and nonsynonymous site-frequency spectra by drawing a random Poisson variates for each entry of the site-frequency spectrum using the observed data as the mean for the variable. We optimized the demographic parameters for each synonymous dataset and then proceeded to optimize the selection parameters under that demographic model. These inferred parameters can be ordered to generate approximate 95% confidence limits that incorporate the uncertainty of both the demographic and selective parameter estimates. Exact 95% limits would require simulating datasets with linkage and simultaneously inferring the demographic and selection parameters. However, the level of linkage in our genome-wide dataset is slight and does not affect our estimates appreciably (Figure 1). Additionally, our algorithm is capable of jointly inferring the demographic and selection parameters, but computationally it is more demanding, and we find that the nonsynonymous site-frequency spectrum adds little power to our ability to infer demography (Figure S3).

**Inference of Mutational Effects across PolyPhen Classes**

For each nonsynonymous SNP in our database, we sought a PolyPhen classification for every transcript in the human genome containing that SNP. Briefly, PolyPhen will classify a SNP as benign, possibly damaging, or probably damaging based on how conserved the site is across a multi-species alignment, functional annotation, and upon the structural consequences of the SNP [26,27]. To assess evolutionary conservation, our sequences were aligned to those in the nrdb95 database (a union of the Swissprot, Swissnew, Trembl, Tremblnew, Genbank, PIR, Wormpep and PDB databases, with sequences of >95% similarity removed; [51]) using BLAST, and conservation was measured using PSIC. Functional annotations were taken from the UniProt database (http://www.pir.uniprot.org/), and structures were taken from PDB or PQS. Excluding SNPs that were unclassifiable or had different PolyPhen classifications for different transcripts, we retained 15,916 benign, 4,199 possibly damaging, and 2,646 probably damaging SNPs. The population-specific site-frequency spectrum for each mutational class was calculated as before, and
selective inference was run assuming class-specific genome-wide mutation rates. To determine what proportion of new non-synonymous mutations PolyPhen classifies as benign, possibly damaging, or probably damaging, we randomly mutated coding sequences from RefSeq 18 using a context-dependent substitution model [45]. We mutated 0.9% of the sites for a total of 46,492 non-synonymous SNPs across 18,657 RefSeq transcripts that were then analyzed using PolyPhen. We found 59.5%, 21.2% and 19.3% of mutations were classified as benign, possibly damaging, and probably damaging, respectively, and used these proportions to calculate the class-specific genome-wide mutation rates. Finally, we also obtained PolyPhen classifications for each human-chimp fixed difference in our dataset.

Supporting Information

Figure S1 Observed versus critical values of the $\chi^2$ goodness of fit test statistic using ms (Hudson 2002) with linkage to determine the test statistic distribution based on 1,000 runs of 22 chromosomes. (A) African expansion model: ms 24 22000 -t 269.1 -r 500 500 -en 0.0064 1 0.3034. (B) European simple bottleneck model: ms 32 22000 -t 308.9 -r 500 500 -en 0.0714 1 0.2692 -en 0.00725 1 0.1898. (C) European complex bottleneck model: ms 32 22000 -t 543.8 -r 500 500 -en 0.02505 1 0.1302 -en 0.02465 1 0.00495 -en 0.0027 1 0.1327.

Found at: doi:10.1371/journal.pgen.1000083.s001 (0.19 MB TIF)

Figure S2 Observed site frequency spectra at silent sites versus expected neutral site frequency spectra under best-fit demographic models. (A) African synonymous sites versus stationary expectation and best-fit expansion model. (B) European synonymous sites versus stationary expectation and best-fit expansion and bottleneck models. Found at: doi:10.1371/journal.pgen.1000083.s002 (0.15 MB TIF)

Figure S3 Likelihood surface plots for demographic inference; $\tau$ is the timing of the expansion event in generations scaled by $2N_{nuc}$. $\omega$ is the ratio of $N_{e, pop}/N_{e, ucc}$. (A) African expansion model inferred solely from synonymous site frequency spectrum. (B) African expansion model inferred by simultaneous inference of demography and selection (assuming gamma distribution) with synonymous and replacement site frequency spectra. Found at: doi:10.1371/journal.pgen.1000083.s003 (6.53 MB TIF)

Figure S4 Likelihood surface plots for selection distribution inference using a gamma model of selective effects: $f(Y = 2N_e \gamma, \alpha, \beta) = -\gamma \alpha^{-1} e^{-\gamma Y/[\beta \Gamma(\gamma)\gamma]}$ for $\gamma < 0$. (A) African ($N_e = 25636$): maxLL at $\alpha = 0.184$, $\beta = 8200$. (B) European ($N_e = 52907$): maxLL at $\alpha = 0.206$, $\beta = 15400$.

Found at: doi:10.1371/journal.pgen.1000083.s004 (6.53 MB TIF)

Table S1 Summary of best-fit demographic models. Found at: doi:10.1371/journal.pgen.1000083.s005 (0.04 MB DOC)

Table S2 Site-frequency spectra and human-chimp fixed difference counts used for inferences in this paper. Found at: doi:10.1371/journal.pgen.1000083.s006 (0.09 MB DOC)

Text S1 Supplementary Material Found at: doi:10.1371/journal.pgen.1000083.s007 (0.03 MB DOC)

Acknowledgments

We dedicate this paper to the memory of our friend and colleague Scott Williamson. We thank Peter Donnelly, Bob Griffiths, and Monty Slatkin for very helpful discussion in the initial stages of this project. We also thank Adi Fleked-Alon for her contribution to an earlier version of protreq, Michele Cargill for her generous assistance and patience with our assembly of the bioinformatic pipeline, and Adam Eyre-Walker and three anonymous reviewers for many helpful comments on earlier drafts of this manuscript.

Author Contributions

Conceived and designed the experiments: AB SW MA JS TW RN AC. Performed the experiments: AB AI JD RH KL MA. Analyzed the data: AB AJD RH KL SS. Contributed reagents/materials/analysis tools: AB SW MA JS TW RN AC. Performed the experiments: AB AI JD RH KL MA. Analyzed the data: AB AJD RH KL SS. Contributed reagents/materials/analysis tools: AB SW AI RH SS RN. Wrote the paper: AB RN AC.

References


Amino Acid Mutations in the Human Genome
Amino Acid Mutations in the Human Genome


