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

Sunjin Moon and Joshua M. Akey

Department of Genome Sciences, University of Washington, Seattle, Washington 98195-5065, USA

A continuing challenge in the analysis of massively large sequencing data sets is quantifying and interpreting non-neutrally evolving mutations. Here, we describe a flexible and robust approach based on the site frequency spectrum to estimate the fraction of deleterious and adaptive variants from large-scale sequencing data sets. We applied our method to approximately 1 million single nucleotide variants (SNVs) identified in high-coverage exome sequences of 6515 individuals. We estimate that the fraction of deleterious nonsynonymous SNVs is higher than previously reported; quantify the effects of genomic context, codon bias, chromatin accessibility, and number of protein–protein interactions on deleterious protein-coding SNVs; and identify pathways and networks that have likely been influenced by positive selection. Furthermore, we show that the fraction of deleterious nonsynonymous SNVs is significantly higher for Mendelian versus complex disease loci and in exons harboring dominant versus recessive Mendelian mutations. In summary, as genome-scale sequencing data accumulate in progressively larger sample sizes, our method will enable increasingly high-resolution inferences into the characteristics and determinants of non-neutral variation.

[Supplemental material is available for this article.]

Copious amounts of exome and whole-genome sequence data have been, and continue to be, generated, yielding massively large catalogs of human genomic variation in geographically diverse populations (Novembre et al. 2008; The 1000 Genomes Project Consortium 2012; Keinan and Clark 2012; Tennessen et al. 2012; Fu et al. 2013). A fundamental challenge in interpreting genome-scale sequencing data derived from increasingly large panels of individuals is identifying and quantifying variants that influence evolutionary fitness. A deeper understanding of deleterious and advantageous mutations would enable insights into the characteristics and determinants of non-neutral variation and have important practical consequences for inferring human demographic history (Fu et al. 2013), informing disease gene mapping studies (Mathieson and McVean 2012; Henn et al. 2015), and clinical genomics (Dewey et al. 2014).

A number of approaches have been pursued to identify or quantify variants that may have functional or fitness effects. For instance, functional prediction methods based on physiochemical properties of nonsynonymous mutations (Kumar et al. 2009; Adzhubei et al. 2010), evolutionary conservation metrics that are applicable to all mutational types (Cooper et al. 2005; Siepel et al. 2006), or statistics that aggregate information across a wide variety of predictive methods are widely used (Kircher et al. 2014). A limitation of functional prediction methods is that they often yield disparate results when applied to the same data set (Fu et al. 2014; Henn et al. 2015), likely reflecting high rates of both false-positive and -negative predictions. Another strategy to quantify non-neutral (primarily deleterious) variation is to explicitly model evolutionary and demographic history from patterns of genetic variation in order to disentangle the effects of selection from confounding evolutionary forces. Although powerful, such models are parameter-rich, and thus inferences are potentially sensitive to model misspecification.

Here, we develop a simple population genetics approach for estimating the fraction of deleterious or adaptive variants in large sequencing data sets. The key advantages of our method are its robustness to a wide range of evolutionary and demographic confounding forces and the ability to quantify patterns of selection in any class of sites of interest. We leverage our method to perform a comprehensive analysis of non-neutral protein-coding variation in exome sequences from 6515 individuals sequenced as part of the Exome Sequencing Project (ESP) (Fu et al. 2013). These analyses reveal new insights into the heterogeneous and context-dependent forces that shape patterns of deleterious nonsynonymous and synonymous variation, characteristics of natural selection that act on disease-associated or -causing genes, and pathways that have experienced adaptive evolution.

Results

A simple nonparametric approach to infer the proportion of sites under selection

The site frequency spectrum (SFS) is a compact summary of genetic variation (Fig. 1A) that contains considerable information about population history (Gutenkunst et al. 2009) and the evolutionary forces that have shaped extant patterns of segregating variation (Akey 2009). For example, purifying selection acting on deleterious alleles results in a skew of the SFS toward rare variation, whereas positive selection acting on advantageous alleles causes a skew of the SFS toward common variation relative to neutral expectations (Fig. 1A). Thus, in principle, the fraction of sites under selection, f, can be inferred by comparing the SFS between a class of variants of interest (which we denote as test sites) to the SFS of putatively neutral variation (which we denote as reference sites). More

Corresponding authors: sunjin@uw.edu, akeyj@uw.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.203059.115.

26:834–843 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/16; www.genome.org

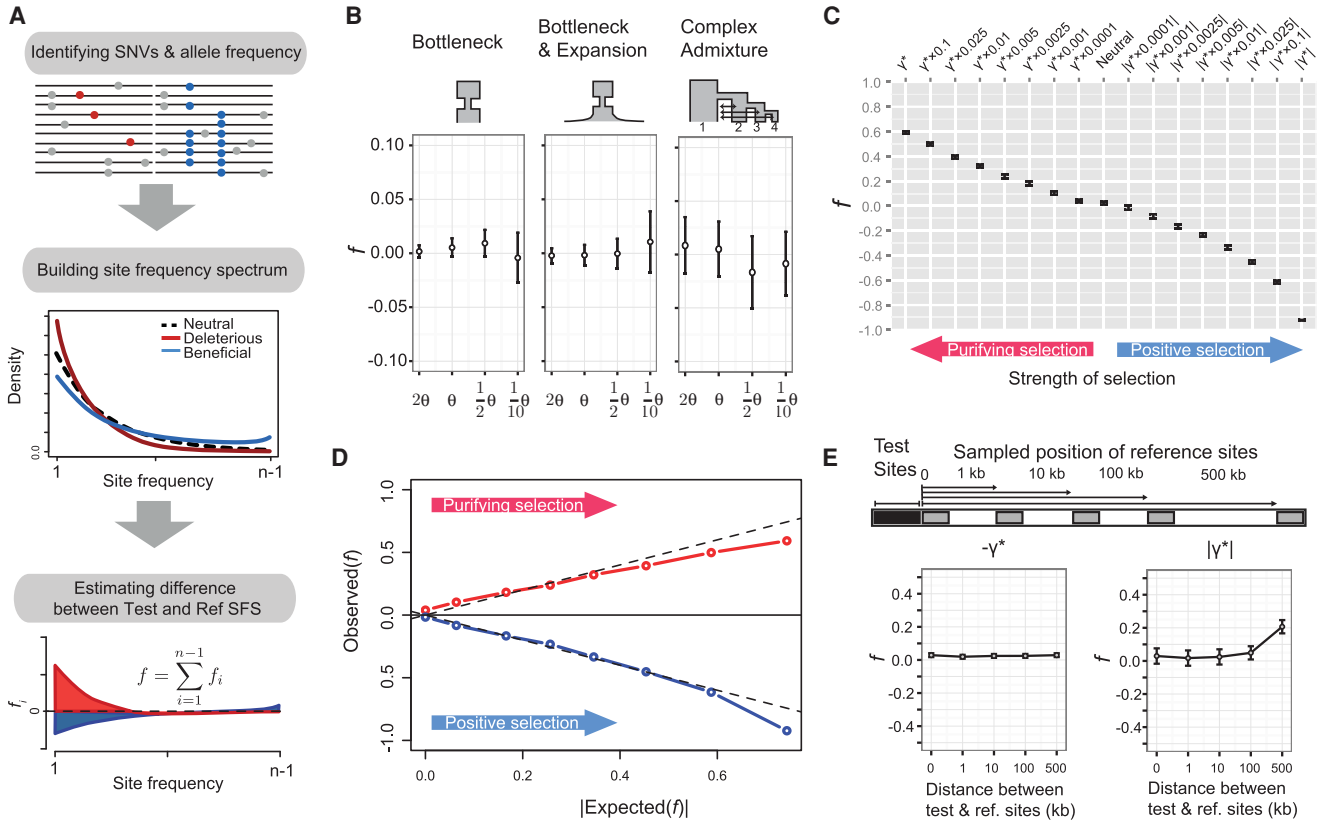

Figure 1. A nonparametric approach for estimating the fraction of SNVs under selection. (*A*) Schematic illustration of the method to estimate *f* by comparing a test SFS to a putatively neutral reference SFS. The fraction of deleterious (red) and advantageous (blue) SNVs is estimated as the scaled difference between test and reference SFS (see Methods). (*B*) Estimates of *f* when test and reference sites have different mutation rates (the test SFS was set to θ) for a different demographic model. The estimate of *f* in the population structure model is from population one. (*C*) Estimates of *f* as a function of strength of selection assuming the same demographic model as in *B*. (γ^*) Baseline selection model (Boyko et al. 2008). (*D*) Comparison between observed and expected estimates of *f* ($|s| > 0.0001$) inferred from the distribution of selection coefficients. (*E*) Estimates of *f* for neutral mutations linked to selected sites (black region) as a function of distance from reference sites (gray regions).

specifically, *f* can be estimated as the difference between a test and reference SFS, summed across all frequency classes (Fig. 1A). With appropriate rescaling, positive and negative values of *f* provide estimates of the fraction of test sites that are under purifying and positive selection, respectively (Fig. 1A). Conceptually, our approach is analogous to methods such as d_N/d_S (Yang 1998) and the McDonald–Kreitman test (McDonald and Kreitman 1991).

We first carefully investigated a number of neutral evolutionary forces that could potentially confound estimates of *f*. To evaluate the effects of heterogeneity in the neutral mutation rate, we simulated data under neutrality where reference sites had a mutation rate of μ and test sites had mutation rates of 0.1μ, 2μ, or 10μ under realistic models of human demographic history (Supplemental Figs. S1, S2). Although mutation rates influence the amount of genetic variation, the shape of the SFS is approximately constant, and thus, estimates of *f* are approximately zero (Supplemental Fig. S2). Furthermore, estimates of *f* are not confounded by demographic perturbations such as population bottlenecks, expansions, and cryptic population structure (Fig. 1B), and therefore, inferences can be made without accurately specifying a demographic model as required by other approaches (Williamson et al. 2004; Boyko et al. 2008; Nielsen et al. 2009). The robustness in estimates of *f* to demographic history is due to the fact that, on average, genetic drift influences both the test and reference class of sites equally. Moreover, estimates of *f* are robust to recombination rate heterogeneity (Supplemental Fig. S3).

Next, to evaluate the accuracy of *f*, we simulated exome sequences under purifying and positive selection. Initially, we assumed the distribution of fitness effects followed a gamma distribution with parameters previously inferred for protein-coding sequences (Boyko et al. 2008) and a realistic model of European demographic history (see Methods) (Tennessen et al. 2012). Under this baseline model of purifying selection (denoted as γ^*), the parameters of the gamma distribution correspond to an average selection coefficient of $s = -3 \times 10^{-2}$. The strength of selection was also systematically varied around the baseline model γ^* (the average *s* ranged from -3×10^{-6} to -3×10^{-2}). There is a clear relationship between *f* and the strength and type of selection (Fig. 1C). For instance, in the baseline model of purifying selection (γ^*), *f* is estimated to be 59% and decreases to 4% when the magnitude of selection is reduced to 10^{-4}. Similarly, in models of positive selection, *f* is estimated to be 92% in the baseline model ($|\gamma^*|$) and 1.6% when the magnitude of selection is reduced by 10^{-4}. Overall, our simulations suggest that *f* can be accurately inferred for sites with selection coefficients ($|s| > 0.0001$) (Fig. 1D). In general, robust estimates of *f* can be made with sample sizes of at least 1000 individuals, with larger sample sizes providing more accuracy (Supplemental Fig. S4). Very strong purifying or positive selection

can result in f being under- or overestimated, respectively, relative to theoretical expectations (Fig. 1D).

Finally, we studied the effects of selection at linked neutral sites (background selection and adaptive hitchhiking) on estimates of f. To this end, we simulated neutrally evolving test sites linked at varying distances to adaptive or deleterious sites (Fig. 1E). Estimates of f were robust to both background selection and hitchhiking when the test and reference sites are separated by <100 kb (Supplemental Figs. S1, S5). Intuitively, these results make sense, as selection will have similar effects on closely linked loci. In hitchhiking models, f can be overestimated when test and reference sites are separated by larger distances (Fig. 1E), and thus, reference sites should be chosen to minimize the effects of strongly advantageous mutations. In summary, these results demonstrate that our simple nonparametric approach for estimating f is accurate and robust to a wide variety of potentially confounding evolutionary forces.

Quantifying the burden of deleterious protein-coding variation

We estimated f in high-coverage exome sequences derived from 4298 European Americans (EAs) and 2217 African Americans (AAs) (Fu et al. 2014). In total, 597,921 coding and 326,065 intronic SNVs are present in these 6515 individuals (446,783 nonsynonymous and 151,138 synonymous) (Supplemental Table S1). As a putatively neutral class of reference sites, we used unconstrained intronic variants proximal to each exon (see Methods). For all protein-coding SNVs, f was 0.374 ± 0.002 (mean ± SE) in EAs and 0.319 ± 0.003 in AAs (Fig. 2A). As expected, the fraction of nonsynonymous SNVs that were deleterious ($f_{EA} = 0.585 \pm 0.002$ and $f_{AA} = 0.524 \pm 0.002$) was considerably higher than synonymous SNVs ($f_{EA} = 0.079 \pm 0.005$ and $f_{AA} = 0.073 \pm 0.004$) (Fig. 2A). The higher estimates of f in EAs compared with AAs are consistent with previous studies (Lohmueller et al. 2008; Fu et al. 2014) and are likely due to differences in demographic history. Note, previous studies based on PolyPhen2 (Adzhubei et al. 2010) or arbitrary conservation thresholds have estimated that less than ~40% of nonsynonymous SNVs are deleterious (Fu et al. 2013); our data driven estimate of f suggests a considerably higher burden of deleterious nonsynonymous SNVs.

Genomic context is a strong determinant of deleterious protein-coding variation at nonsynonymous sites

Next, to better understand how patterns of selection are influenced by genomic context, we categorized synonymous and nonsynonymous variants according to whether the ancestral allele at each SNV occurred in a CpG site, a potential site of GC-biased gene conversion (gBGC), or a non-CpG and non-gBGC site (NCB; see Methods). We constructed the reference and test SFS to have the same mutational types to mitigate influences of mutation rate heterogeneity (although simulations demonstrate this effect is likely to be small) (Fig. 2B). Indeed, estimates of f were robust to the class of reference sites used, except in comparisons that involved gBGC sites (Supplemental Figs. S6, S7), likely because gBGC increases the rate of fixation and skews the SFS toward high frequencies. Thus, matching genomic context between test and reference sites is important when disentangling the effects of selection from other evolutionary forces.

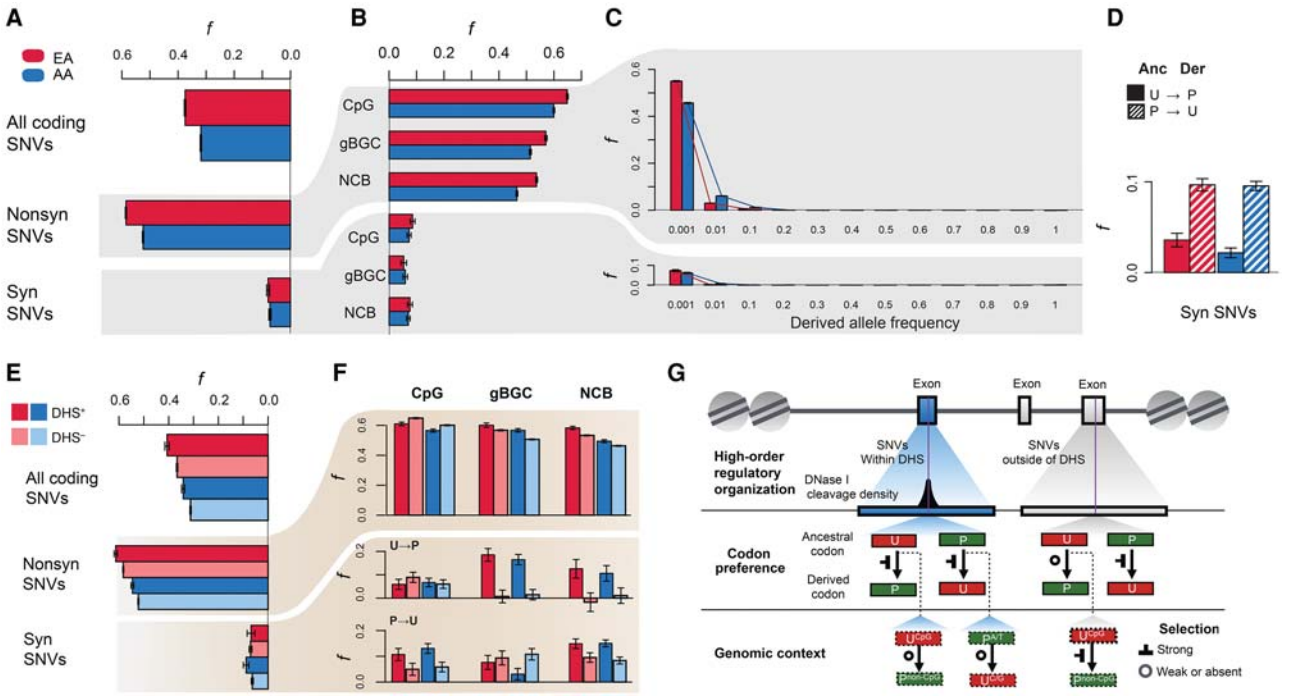

Figure 2. Characteristics of deleterious protein-coding SNVs. Estimates of f for (*A*) all protein-coding SNVs, all nonsynonymous SNVs, and all synonymous SNVs in EAs (red) and AAs (blue). (*B*) Estimates of f for SNVs as a function of genomic context. (*C*) Estimates of f as a function of derived allele frequency. (*D*) Effect of codon bias on estimates of f at synonymous sites. (*E*) Estimates of f for SNVs inside (+; darker shade) or outside (−; lighter shade) of DHSs. (*F*) Decomposing the effects of genomic context, codon preference, and DHSs on estimates of f. (*G*) Schematic summary of context-dependent patterns of deleterious synonymous SNVs (purple, vertical lines) as a function of regulatory context (first row), codon preference (second row), and genomic context (third row). Note that preferred changes (U → P) at CpG sites and unpreferred changes (P → U) of gBGC within the DHSs show less constraint, whereas preferred changes (U → P) at CpG sites outside of DHSs exhibit stronger constraint.

Genomic context has limited effects at synonymous sites, with estimates of f ranging from approximately 0.053 to 0.085 in EAs and 0.059 to 0.073 in AAs (Fig. 2B). At nonsynonymous sites, however, genomic context has a profound influence on estimates of f. Specifically, the highest fraction of deleterious variants occurs in CpG sites ($f_{EA} = 0.649 \pm 0.003$ and $f_{AA} = 0.603 \pm 0.004$), whereas NCB sites exhibit the lowest fraction of deleterious nonsynonymous variants ($f_{EA} = 0.537 \pm 0.004$ and $f_{AA} = 0.465 \pm 0.003$) (Fig. 2B). The significantly higher estimates of f in CpG compared with gBGC or NCB sites (Wilcoxon text, $P < 10^{-16}$) (Supplemental Fig. S8) suggests that CpG hypermutability (Shen et al. 1992) is a potent source of deleterious nonsynonymous mutations. Consistent with this hypothesis, nonsynonymous SNVs in CpG sites have significantly higher average PolyPhen2 (Adzhubei et al. 2010) and Grantham (1974) scores than non-CpG sites (mean PolyPhen2 scores of 0.687 ± 0.001 and 0.552 ± 0.001, respectively; Mann-Whitney U test, $P < 10^{-16}$; mean Grantham scores of 74.8 ± 0.1 and 69.9 ± 0.1, respectively; $P < 10^{-16}$). Note, unless otherwise stated, all P-values reported below are from Mann-Whitney U tests.

To explore the stability of f estimates, we repeated all analyses using four different sets of reference sites (Supplemental Fig. S9). Overall, estimates of f were extremely robust. For example, in EAs f ranged from 0.656 to 0.675 for nonsynonymous CpG sites across different sets of reference sites. We also estimated f using only conserved intronic sites as the reference SFS. As expected, f was reduced ~7% at nonsynonymous sites and ~14% at synonymous sites (Supplemental Fig. S9), likely due to the presence of deleterious variants in the reference SFS. Thus, f may be underestimated in empirical data sets, as it is difficult to unambiguously define sites evolving under strict neutrality, although this effect is likely to be modest when reference sites are carefully chosen.

Rare variants are highly enriched for deleterious alleles in EA

We estimated the fraction of deleterious variants as a function of derived allele frequency (Fig. 2C). Overall, 94% of deleterious SNVs in EAs and 87% of deleterious SNVs in AAs were rare (DAF < 0.001), consistent with previous studies (Fu et al. 2013). Notably, estimates of f for singletons (derived alleles that appear once in the sample) were $f_{EA} = 0.427$ and $f_{AA} = 0.349$ for nonsynonymous SNVs and $f_{EA} = 0.059$ and $f_{AA} = 0.046$ for synonymous SNVs. To more directly compare f between EAs and AAs, we sampled an equal number of chromosomes from each population (Supplemental Fig. S10). In all mutational classes, EAs have significantly higher estimates of f compared with AAs ($P < 10^{-16}$), consistent with previous observations that the out-of-Africa bottleneck resulted in proportionally more deleterious variation in EAs compared with AAs (Lohmueller et al. 2008; Fu et al. 2014).

Influence of codon bias and chromatin accessibility on the burden of deleterious SNVs

Codon bias refers to the differential use of synonymous codons and has been found in a wide variety of organisms (Novoa et al. 2012), although the amount of codon bias and the evolutionary forces influencing it in humans remains unclear (Kotlar and Lavner 2006; Yang and Nielsen 2008). To investigate the evolutionary dynamics of codon bias in humans, we focused on derived mutations that result in a change from a preferred to unpreferred codon (P → U) or a change from an unpreferred to preferred codon (U → P) (Fig. 2D). Estimates of f for SNVs with P → U ($f^{EA} = 0.097 \pm 0.006$, $f^{AA} = 0.093 \pm 0.005$) codon changes were significantly higher than SNVs with U → P codon changes ($f^{EA} = 0.039 \pm 0.008$, $f^{AA} = 0.020 \pm 0.006$) ($p^{EA} = 2.7 \times 10^{-7}$, $P^{AA} = 1.0 \times 10^{-14}$) (Fig. 2D), suggesting stronger constraint on P → U SNVs.

Next, we estimated f for variants in DNase I hypersensitive sites (DHSs), which delimit regions of open chromatin, and contrasted it to estimates of f for variants outside of DHSs. Overall, f is significantly higher for variants in DHSs compared with variants outside of DHSs in both the EA and AA samples ($P^{EA} < 10^{-16}$, $P^{AA} = 1.7 \times 10^{-15}$) (Fig. 2E), suggesting higher levels of constraint in protein-coding regions that may also encode regulatory information (Stergachis et al. 2013). This pattern was particularly strong at nonsynonymous variants (Fig. 2E), although the relationship between f from SNVs within and outside of DHSs was strongly influenced by genomic context (Fig. 2F). For instance, estimates of f were significantly higher within compared with outside of DHSs for gBGC ($P^{EA} = 6.6 \times 10^{-6}$, $P^{AA} = 8.2 \times 10^{-7}$) and NCB ($P^{EA} = 2.4 \times 10^{-6}$, $P^{AA} = 0.006231$) sites, whereas for CpG sites they were significantly lower ($P^{EA,AA} < 10^{-16}$) (Fig. 2F).

Surprisingly, estimates of f for synonymous variants within and outside of DHSs were similar, particularly in EAs (Fig. 2E). As heterogeneous and context-dependent forces likely influence patterns of evolutionary constraint, we decomposed the effects of codon bias, chromatin accessibility, and genomic context on estimates of f (Fig. 2F). We found distinct patterns of selection when simultaneously accounting for all of these factors, which are not readily apparent when considering any factor in isolation (Fig. 2F). For example, the equivocal estimates of f for synonymous variants within and outside of DHSs arise from the complex, and opposing, patterns caused by codon bias and genomic context. Specifically, estimates of f at U → P synonymous variants in gBGC sites are significantly higher within compared with outside of DHSs ($P^{EA} = 8.5 \times 10^{-6}$, $P^{AA} = 1.0 \times 10^{-6}$) (Fig. 2F), whereas they are similar at CpG sites. Conversely, estimates of f at P → U synonymous variants in CpG sites are higher within compared with outside of DHSs (Fig. 2F). It is interesting to note that differences between f within and outside of DHSs have the same patterns in EAs and AAs, except for P → U gBGC synonymous sites (Fig. 2F). We hypothesize this observation may be due to the interaction of recombination rate heterogeneity between populations, strength of gBGC, and genomic contexts (Galtier et al. 2009; Glemin et al. 2015), but additional work is necessary to fully interpret patterns of deleterious variation at P → U gBGC synonymous sites. A schematic summary of the context-dependent effects observed in estimates of f for synonymous SNVs is shown in Figure 2G.

Selection at the center and periphery of protein–protein interaction networks

To better understand how interactions between proteins influence characteristics of deleterious variation, we constructed a protein–protein interaction (PPI) network consisting of 6700 protein-coding genes that harbor variants in the ESP data (Fig. 3A). Consistent with previous studies, the topology of the network is approximately scale-free (Barabási and Oltvai 2004; Stelzl et al. 2005) such that most proteins have relatively few interactions, although some proteins have a large number of interactions (Fig. 3B). There is a significant ($P = 0.009$) linear relationship between estimates of f at nonsynonymous sites and the number of PPIs, with a 10-fold increase in interactions associated with a 3.3% increase of f (Fig. 3C). Moreover, the difference between estimates of f at nonsynonymous sites for nodes in the center and periphery of the network is about 0.11 ($f_{EA} = 0.66$ and $f_{AA} = 0.62$ in the center

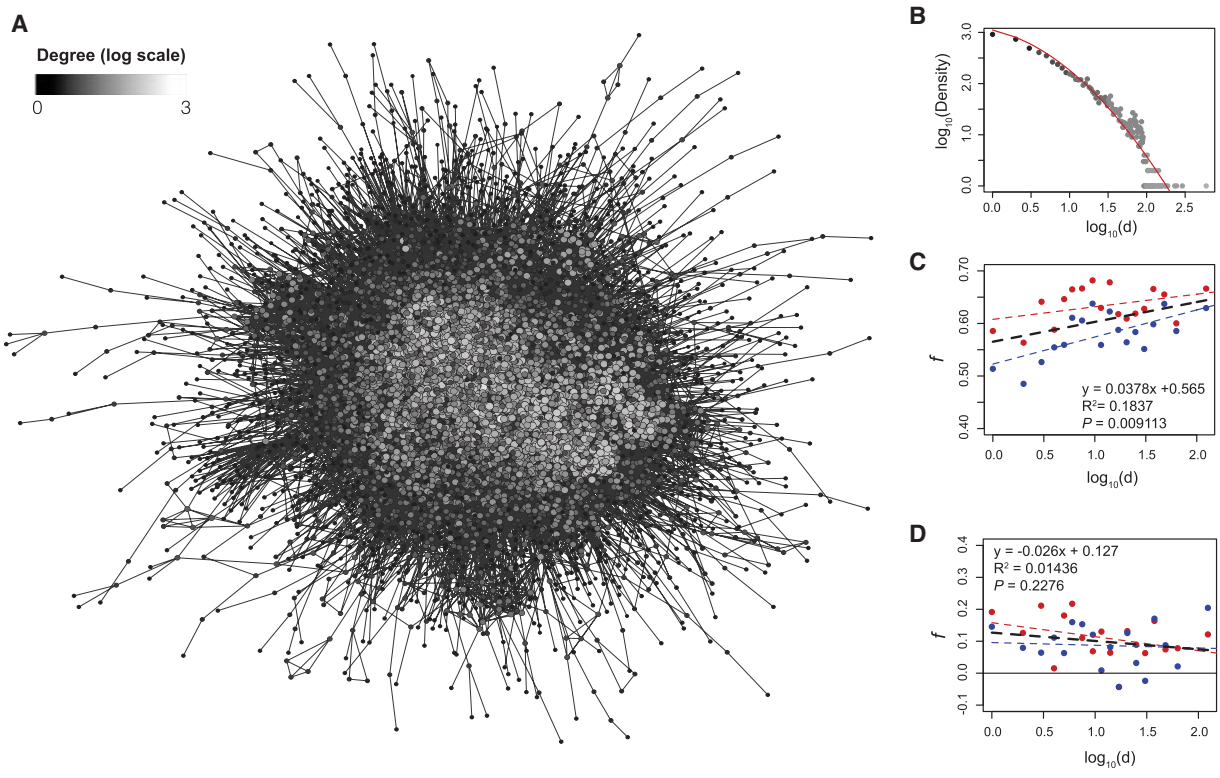

Figure 3. Higher burden of nonsynonymous SNVs for proteins at the center versus periphery of protein–protein interaction networks. (*A*) PPI network of 6700 protein-coding genes with 56,728 significant connections (STRING score >0.7). Each protein is a node and is shaded by the number of interactions (degree). (*B*) Relationship between the average number of interacting proteins (*x*-axis) and the degree of connectivity (*y*-axis) for the PPI network. (*C*) Relationship between the average number of interacting protein neighbors on estimates of *f* at nonsynonymous SNVs, showing significantly increased purifying selection on genes located at center rather than periphery of the PPI network. (*D*) No significant relationship between the average number of interacting proteins on estimates of *f* at synonymous sites.

and $f_{EA} = 0.58$ and $f_{AA} = 0.51$ in the periphery). When separating variants into genomic context, the effect of interactions becomes more pronounced (Supplemental Fig. S11); differences in *f* can be as high as 40% between nodes in the center and periphery of the network at CpG sites. Interestingly, there is no significant relationship between estimates of *f* and the number of interactions at synonymous sites (Fig. 3D).

Selection on nonsynonymous SNVs in disease-related genes

We also studied patterns of selection across disease related genes. Specifically, we first estimated *f* for nonsynonymous SNVs in genes annotated as "Essential" (731 genes) (Bult et al. 2013), "Mendelian" disorders (Amberger et al. 2015) (2622 genes), "Complex" traits (Becker et al. 2004) (1690 genes), and "Other" (2472 genes). Estimates of *f* were significantly different across categories ($P < 0.01$ for all comparison in EAs and AAs), with Essential genes showing the most constraint ($f_{EA} = 0.644 \pm 0.008$, $f_{AA} = 0.610 \pm 0.008$), followed by Mendelian disease genes ($f_{EA} = 0.607 \pm 0.006$, $f_{AA} = 0.527 \pm 0.006$), Complex disease genes ($f_{EA} = 0.528 \pm 0.008$, $f_{AA} = 0.467 \pm 0.006$), and Other genes ($f_{EA} = 0.453 \pm 0.009$, $f_{AA} = 0.407 \pm 0.007$) (Fig. 4A).

Next, we identified exons that harbor clinically significant mutations as defined by ClinVar (Landrum et al. 2014), categorized these exons based on the mutations' reported mode of inheritance (Dominant, Recessive, or Other), and estimated *f* from nonsynonymous SNVs found within these exons (Fig. 4B). As expected, the average number of nonsynonymous SNVs per kb in ClinVar exons is higher in the ESP data compared with SNV density estimated directly from ClinVar (Fig. 4B) given the larger sample size. Strikingly, estimates of *f* in exons associated with autosomal-dominant diseases ($f_{EA} = 0.807 \pm 0.007$, $f_{AA} = 0.704 \pm 0.008$) were significantly higher than estimates of *f* from exons harboring autosomal-recessive disease ($f_{EA} = 0.559 \pm 0.020$, $f_{AA} = 0.502 \pm 0.016$; $P < 10^{-16}$) (Fig. 4B). Indeed, estimates of *f* from exons with dominant diseases are significantly higher ($P^{EA,AA} < 10^{-16}$), whereas estimates of *f* from exons with recessive diseases are significantly lower ($P^{EA} = 0.017$, $P^{AA} < 0.031$), compared with *f* from nonsynonymous SNVs in exons associated with other ClinVar diseases ($f_{EA} = 0.626 \pm 0.018$, $f_{AA} = 0.552 \pm 0.018$) (Fig. 4B). It is interesting to note the ~30% reduction of *f* in recessive compared with dominant exons may reflect the proportion of potentially deleterious nonsynonymous mutations in heterozygous form with no visible deleterious phenotype in recessive exons.

Estimates of *f* facilitate probabilistic interpretations of *C*-scores

Recently, an approach to assess the pathogenicity of individual variants was developed that integrates a large number of heterogeneous data types into a single metric denoted as a *C*-score (Kircher et al. 2014). To date, interpreting *C*-scores has relied upon arbitrary empirical thresholds. To provide a more probabilistic interpretation of *C*-scores, we calculated *f* for ESP variants as a function of scaled *C*-score bin for nonsynonymous (Fig. 5A) and synonymous

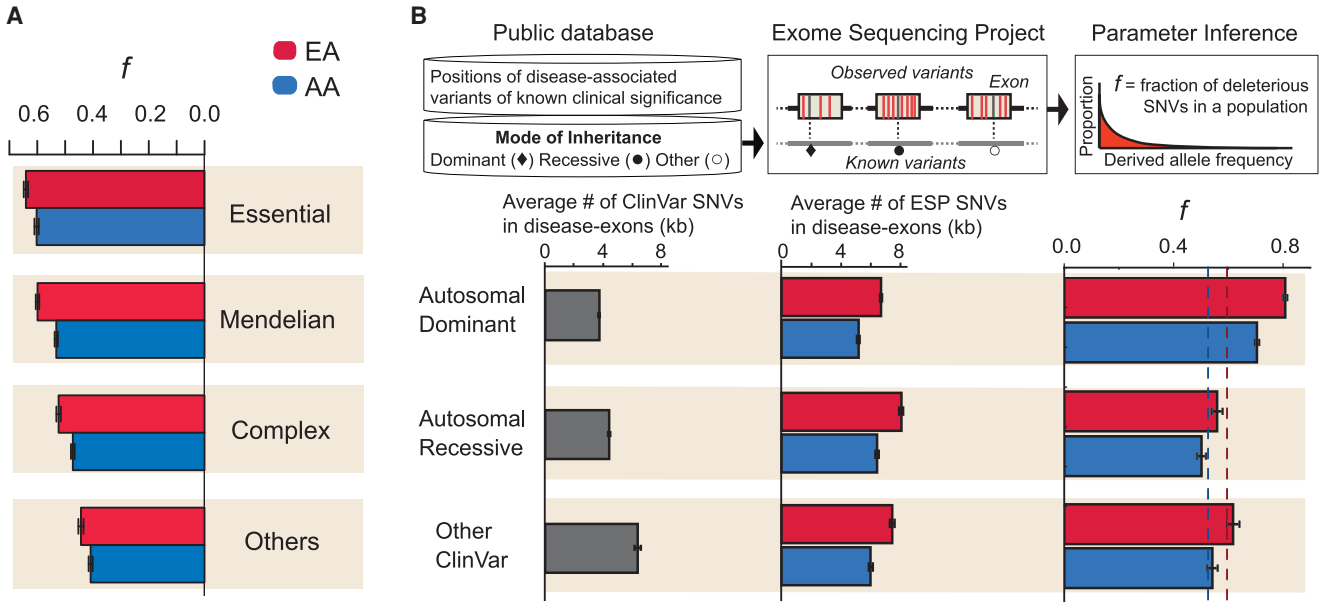

Figure 4. Intensity of purifying selection on nonsynonymous SNVs in disease related genes. (*A*) Estimates of *f* in genes designated as Essential, Mendelian, Complex, or Others (for details, see text). (*B*) Estimates of *f* from SNVs in exons with reported disease causing mutations as a function of mode of inheritance. The *top* panel provides a schematic illustration of the analysis, including intersection between exons containing ClinVar reported mutations and nonsynonymous SNVs in ESP. The plots *below* show the number of nonsynonymous SNVs/kb estimated from ClinVar and ESP, as well as estimates of *f* as a function of mode of inheritance. Dashed lines denote the average *f* for nonsynonymous SNVs in EAs (red) and AAs (blue).

(Fig. 5B) SNVs. For nonsynonymous SNVs, there was a significant relationship between *f* and scaled *C*-score ($R^2_{EA} = 0.94$, $P = 6.3 \times 10^{-120}$; $R^2_{AA} = 0.95$, $P = 2.6 \times 10^{-129}$) (Fig. 5A), with $f > 0.90$ for scaled *C*-scores greater than about 18 (i.e., we estimate 90% of nonsynonymous SNVs with a scaled *C*-score ≥18 are deleterious). Importantly, the relationship between *f* and scaled *C*-scores was similar in both EAs and AAs (Fig. 5A). Note, estimates of *f* for the lowest scaled *C*-score bin are about 0.10, suggesting that a small fraction of nonsynonymous SNVs that possess little evidence of being pathogenic by even sophisticated prediction algorithms may actually be deleterious. The relationship between *f* and scaled *C*-scores for synonymous SNVs was considerably more modest ($R^2_{EA} = 0.60$, $P = 8.5 \times 10^{-12}$; $R^2_{AA} = 0.63$, $P = 6.3 \times 10^{-16}$) (Fig. 5B) and varied around zero for *C*-scores below five. In summary, estimates of *f* can help guide the interpretation and selection of appropriate *C*-score thresholds, particularly for nonsynonymous SNVs.

Natural selection on pathways

We calculated estimates of *f* in categories of genes as defined by Gene Ontology (GO) (The Gene Ontology Consortium 2015), KEGG pathways (Kanehisa et al. 2014), and Reactome pathways (Croft et al. 2014) in EAs and AAs (Fig. 6; Supplemental Table S2–S4). As expected, the majority of categories and pathways exhibit purifying selection, although there is considerable heterogeneity in levels of constraint (Fig. 6). The most constrained processes and pathways are in general related to gene ensembles that participate in core cellular functions; for example, the largest estimate of *f* in KEGG pathways is for "basal transcription factors." Although estimates of *f* were largely similar between EAs and AAs, a number of categories and pathways exhibited differences in the intensity of purifying selection (Fig. 6). For instance, the GO category "NAD metabolic process" ($f_{EA} = 0.18$ and $f_{AA} = 0.74$) and Reactome pathway "Notch-HLH transcription pathway" ($f_{EA} = 0.02$ and $f_{AA} = 0.64$) had markedly different estimates of *f* among populations. Finally, several categories exhibited evidence of positive selection ($f < 0$) in either one population or both populations (Fig. 6; Supplemental Table S2–S4). Examples of pathways with evidence of positive selection in both EAs and AAs include the KEGG pathway "asthma" ($f_{EA} = -0.93$ and $f_{AA} = -0.28$), GO category "positive regulation of innate immune response" ($f_{EA} = -0.89$ and $f_{AA} = -0.35$), and Reactome pathway "defensins" ($f_{EA} = -0.74$ and $f_{AA} = -0.84$). We caution that strong positive selection can lead to biased estimates of *f* (Fig. 1E). Furthermore, we note that the interpretation of pathway data is complicated by the fact that genes are often assigned to multiple categories and that current pathways are simplified representations of complex biological processes. Nonetheless, these results demonstrate that estimates of *f* yield insight into the tempo and mode of selection acting in aggregate across ensembles of genes.

Discussion

We developed a simple nonparametric method to estimate the fraction of non-neutral SNVs in large sequencing data sets and showed that it is robust to many potential confounding demographic and evolutionary forces. A particularly powerful feature of our approach is its flexibility, which we leverage to comprehensively study characteristics and patterns of non-neutral protein-coding variation. Our results provide important new insights into the heterogeneous and highly context-dependent effects that shape patterns of deleterious protein-coding variation. For example, recent studies (Stergachis et al. 2013; Xing and He 2015) have come to disparate conclusions on whether exonic DHSs provide an additional level of constraint on protein-coding sequences. Our results show that this straightforward question is complicated by the multitude of context-dependent factors that influence

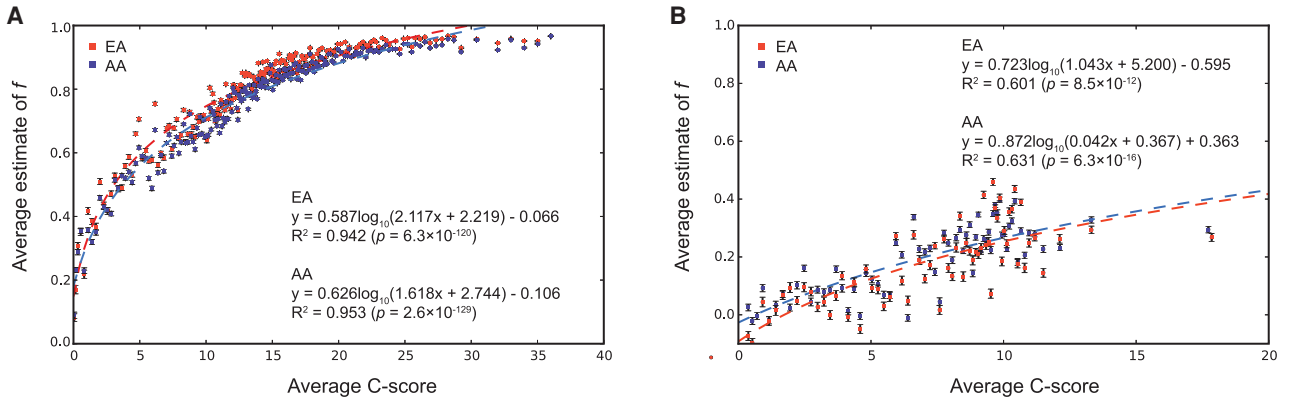

Figure 5. Relationship between predicted pathogenicity of SNVs and estimate of *f* in populations. The average estimate of *f* on nonsynonymous (*A*) and synonymous (*B*) SNVs was computed on each bootstrapped subset of SNVs in each bin by using a quantile-binning approach, such that each bin had the same number of SNVs. Error bars, SEM scaled *C*-score (*x*-axis) and mean estimate of *f* (*y*-axis). The red (EA) and blue (AA) lines represent the best fit curves to the data.

protein-coding variation, but when properly taken into account, regulatory DNA in coding regions does impose additional constraints (Fig. 2E,F), although the magnitude of such effects are modest. Moreover, we found that estimates of *f* were significantly higher for nonsynonymous SNVs in exons that harbor dominant versus recessive diseases. This observation suggests mutations that occur in a particular exon may be more likely to have a similar mode of inheritance. Furthermore, we showed a strong quantitative relationship between estimates of *f* and *C*-scores, particularly for nonsynonymous SNVs (Fig. 5), which enable variants with a particular *C*-score to be interpreted in a probabilistic framework and should be of considerable utility in evolutionary and clinical genomics studies.

It is important to note that our method has several limitations (see also Supplemental Material). For example, accurate estimates of *f* require a reference SFS composed of neutrally evolving sites. In practice, unambiguously identifying neutral sites is challenging, although a number of functional and evolutionary genomics resources can help inform what sites are most likely to be free of selective constraint. Furthermore, although a number of assumptions are required for model-based estimates of the fraction of deleterious variants, they have some advantages compared with our nonparametric approach. For example, they provide more direct estimates on the distribution of fitness effects (Racimo and Schraiber 2014) and other parameters that may be of interest, and thus, our method is complimentary to existing approaches. Of particular interest, fitCons (Gulko et al. 2015) was recently developed to provide an estimate for fitness consequences of mutations at each site in the genome and leverages both polymorphism within and divergence between species, as well as functional genomics data. Thus, fitCons captures the effects of selection acting over longer time-scales compared with our method, which focuses on the effects of more recent selection acting on a class of sites of interest.

Finally, another limitation of our method is interpreting estimates of *f* when test sites are composed of a mixture of deleterious and advantageous mutations. Indeed, we simulated test SFSs composed of varying fractions of advantageous, deleterious, and neutral mutations (Supplemental Fig. S12) and found that estimates of *f* reflect the net effect of selection. In general, deleterious mutations vastly outnumber advantageous mutations, and as a consequence, the power of our method to detect positive selection may be low when aggregating over large numbers of sites (i.e., the signature of positive selection would be attenuated in the presence of deleterious variation). Nonetheless, as sample sizes increase, simulations suggest that future estimates of *f* could be made

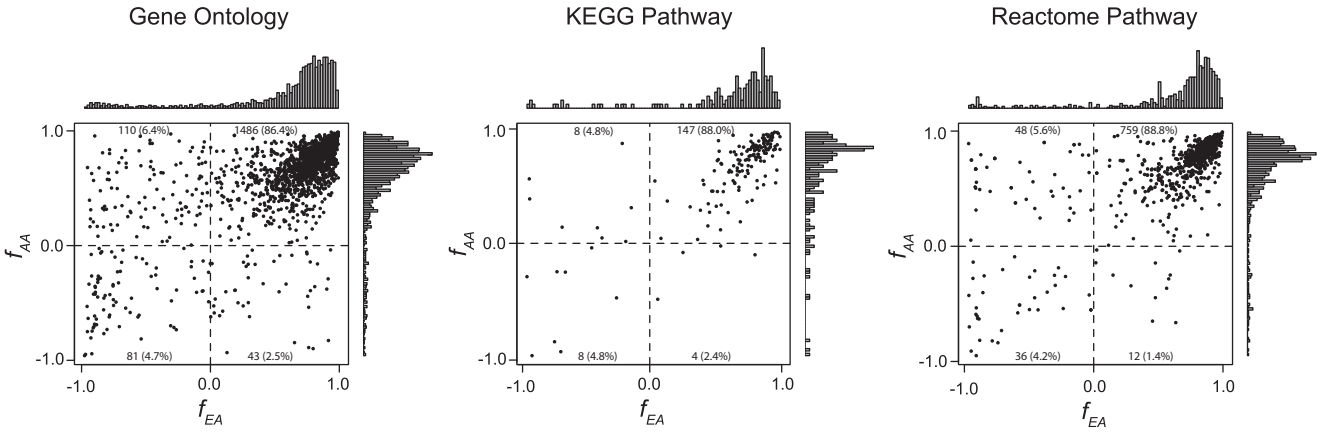

Figure 6. Natural selection on pathways. Estimates of *f* in EAs and AAs were calculated for nonsynonymous SNVs for genes across all categories in Gene Ontology, KEGG pathways, and Reactome pathways. The number (and proportion) of categories in each of the four quadrants is shown in each panel.

at individual genes or exons (Supplemental Fig. S4), which would facilitate the construction of more homogeneous test SFSs and mitigate these interpretational issues.

In conclusion, as whole-genome sequences supplant exomes, and sample sizes move from thousands to hundreds of thousands of individuals, our approach will become an increasingly useful and powerful tool to comprehensively investigate the characteristics and determinants of evolutionarily significant variation.

Methods

The model

Given a set of n DNA sequences, derived variants can be described as a vector of the number of variable sites at frequency of i/n in the sample, where i is the number of observed derived alleles. Thus, the SFS can be written as $\eta = \eta_i$ ($i = 1, 2, \ldots, n-1$). The goal of our approach is to estimate the fraction, f, of non-neutral SNVs in a class of sites of interest (test sites) by comparing the SFS between test sites (η_{test}) to a SFS composed of putatively neutral reference sites (η_{ref}). However, because of differences in effect sequence length and number of SNVs between η_{test} and η_{ref}, it is not possible to compare them directly. Therefore, we first use estimators of the population mutation rate, $\theta = 4N_e\mu L$ (where N_e is the effective population size, μ is the mutation rate/base pair, and L is the length in base pairs of sequence), to scale η_{ref}. As deleterious SNVs are less likely than neutral variants to drift to high frequencies, we us θ_π, which puts more weight on intermediate-frequency variants (Tajima 1983), as an estimator for θ. We calculate the expected value of $\hat{\theta}_\pi$ from the observed SFS as

$$\hat{\theta}_\pi(\eta) = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} i(n-i)\eta_i = 4N_e\mu L.$$

Under a variety of demographic models (Fig. 1), we showed that the total number of observed SNVs, $\sum \eta$, is proportional to μL in a way that the SFS is shifted up or down without changing its shape. This property enables the amount of low-frequency sites to be inferred from the amount of intermediate-frequency sites in two sets of SFSs when μ is unknown and L varies considerably. Next, we calculate the ratio of θ_π estimated from the test (η_{test}) and reference SFS (η_{ref}):

$$\alpha_1 = \frac{\hat{\theta}_\pi(\eta_{test})}{\hat{\theta}_\pi(\eta_{ref})} = \frac{\mu_{test}L_{test}}{\mu_{ref}L_{ref}}.$$

By using α_1, which is the ratio of μL weighted by putatively neutral intermediate-frequency variants between test and reference sites, we obtain the scaled SFS of test sites, $\eta_{ref^*} = \alpha_1\eta_{ref} = \alpha_1\eta_i$ ($i = 1, 2, \ldots, n-1$). Although we could now directly calculate f, we expect most deleterious SNVs to be rare. Therefore, for inferring the fraction of deleterious SNVs in test sites, we use θ_W, which places more weight on low-frequency sites (Watterson 1975):

$$\hat{\theta}_W(\eta) = 1/h_n \sum_{i=1}^{n-1} \eta_i,$$

where $h_n = \sum_{k=1}^{n-1}(1/k)$. We obtain the estimate of f in test sites by removing the fraction of neutral variants inferred from the scaled reference SFS (η_{ref^*}):

$$f = 1 - \frac{\hat{\theta}_W(\eta_{ref*})}{\hat{\theta}_W(\eta_{test})}.$$

Under the null hypothesis that all test sites are neutrally evolving, η_{test} is equal to that of η_{ref^*} (and thus $f \approx 0$). Note, in practice we found estimates of f to be stable regardless of what particular estimator of θ is used (see Supplemental Material).

Estimates of $f < 0$ indicate an excess of common variants in test sites, which may occur if some test sites are subject to positive selection. To estimate the fraction of sites that are under positive selection, we obtain a new scale factor, α_2, for the reference SFS based on θ_W,

$$\alpha_2 = \frac{\hat{\theta}_W(\eta_{test})}{\hat{\theta}_W(\eta_{ref})} = \frac{\mu_{test}L_{test}}{\mu_{ref}L_{ref}}.$$

By using α_2, we have a new $\eta_{ref^*} = \alpha_2\eta_{ref} = \alpha_2\eta_i$ ($i = 1, 2, \ldots, n-1$) to subtract neutral variation at intermediate-frequency sites from η_{test}. It is straightforward to compute the fraction of putatively adaptive variants based on θ_π:

$$f = 1 - \frac{\hat{\theta}_\pi(\eta_{ref*})}{\hat{\theta}_\pi(\eta_{test})}.$$

Note, we use a negative sign to distinguish it from estimates of the fraction of deleterious SNVs.

Simulations

We tested the accuracy and robustness of estimating f under complex demographic scenarios (Tennessen et al. 2012; Fu et al. 2013; Gazave et al. 2013) and a variety of selection regimes. For each demographic scenario (see Supplemental Note), we simulated DNA sequences consisting of coding and noncoding regions. Only nonsynonymous sites in coding regions were under selection with differing selection coefficients, s, a population-scaled selection coefficient $\gamma = 2N_e s$, where γ was drawn from Gamma distribution (Γ). We set the shape (α) and the rate (β) parameters as described by Boyko et al. (2008), where $\alpha^* = 0.206$, $\beta^* = 1/2740$, and $N_e = 10{,}000$ (Gazave et al. 2013), for a baseline model of selection. Γ^* was shifted to have average selection coefficients (α/β) ranging from 1/10,000 to 1/10 of that of the baseline model by multiplying coefficients (10, 40, 100, 200, 400, 1000, 10,000) to the rate parameter (β) with the shape (α) parameter of baseline model of selection. We specified negative and positive γ for negative and positive selection coefficients, respectively. We simulated sequence samples, comparable in size to empirical exome data, consisting of 300 bp of coding and 100 bp of noncoding sequence at each locus. Coding regions were composed of nonsynonymous and synonymous sites, as well as nonsynonymous sites under selection. For evaluating the effect of hitchhiking and background selection on the estimate of f, we specified the distance between coding region and noncoding region to be 0, 1, 10, 100, and 500 kb. Unless otherwise noted, for each parameter combination we simulated 100 replicates of an aggregate set of 1000 loci. Simulations were carried out with varying sample sizes of $n = 10$, 100, 1000, and 10,000 sequences. We performed forward-time simulations implemented in SFS_CODE program (Hernandez 2008).

Exome sequencing data

We analyzed high-coverage exome sequencing data from 6515 individuals that were sequenced as part of the ESP (Fu et al. 2013). QC and data filtering were performed as previously described. Additionally, we defined ancestral alleles based on ancestral genome sequence for *Homo sapiens* (GRCh37) that can be downloaded from Ensembl (ftp://ftp.ensembl.org/pub/release-65/fasta/ancestral_alleles/). We considered only sites showing ancestral allele with high-confidence calls, i.e., ancestral state supported by more than two sequences.

Sequence contexts of SNVs

SNVs were classified into coding and noncoding sites, and coding SNVs were classified into nonsynonymous and synonymous mutations. Nonsynonymous SNVs were counted as a derived allele that resulted in a change of amino acid sequence. Synonymous mutations were defined as changes at the third position of four- and sixfold degenerate sites without impact on the amino acid sequence. Reference noncoding mutations were defined as any mutations located within intronic regions and within 50 bp of 5′-upstream and 3′-downstream regions of exon boundaries. The longest transcript was chosen to identify exon–intron boundaries when there were more than two alternative splicing forms for a gene.

SNVs were classified into three classes based on genomic context: CpG, gBGC, and NCB classes. Specifically, the CpG class was defined as SNVs that occurred in ancestral CpG sites (CpG → NpG or CpN). gBGC sites were defined as SNVs that were weak-to-strong substitutions (AT → GC). All other SNVs were classified as NCB, including A ↔ T, G ↔ C, and GC → AT changes.

Finally, we classified synonymous SNVs into "preferred" or "unpreferred" changes based on the abundance of tRNA in the genome (Novoa et al. 2012). A "preferred" change is defined when the derived allele leads to the use of a tRNA whose anticodon frequency is higher than that of the ancestral allele's tRNA anticodon. Similarly, an "unpreferred" change is defined as cases where the derived allele leads to the use of a tRNA whose anticodon frequency is lower than that of the ancestral allele's tRNA anticodon. Mutations that have no effect on the abundance of tRNA for the ancestral and derived codons were excluded.

Constraint categories of SNVs

SNVs were classified into four categories according to the information available from external experiments: (1) evolutionary conservation, (2) DNase I hypersensitivity analysis, (3) hominid-specific selection, and (4) recombination rate variation. SNVs were classified into different levels of the evolutionary conservation based on PhyloP scores (Cooper et al. 2005) of 99 vertebrate genomes with human genome. SNVs were classified into regulatory potentials by identifying any SNVs overlapped with DHSs identified in 81 human cell types (Stergachis et al. 2013). SNVs were classified into 20 groups according to levels of local recombination rates. More details can be found in the Supplemental Material.

PPI and GO analysis

Genes harboring SNVs were integrated into PPI networks by using the STRING v9 database (Franceschini et al. 2013). Only high-confidence predictions (STRING score >0.7) were included. We focused on the largest cluster of the PPI network that consisted of 56,728 nonredundant interactions among 6700 gene products. Genes were classified into 20 groups according to the degree of interactions, and each group includes approximately similar numbers of genes using the quantile binning procedure described above. The PPI network was visualized using Cytoscape (Smoot et al. 2011).

Genes harboring SNVs were grouped into categories of GO biological processes retrieved from BioMart (*Homo sapiens* genes version GRCh37.p13 in Ensemble Genes version 75). We considered all GO terms that had at least 10 genes, retaining a total of 1719 GO biological process terms. Because many terms were "part-of" relationships, we applied FDR, a less-conservative method than Bonferroni, to control for multiple testing.

Complex diseases association

We classified genes into Mendelian disease, Essential genes, Complex disease, and other as previously described (Fu et al. 2013). Briefly, the list of genes associated with Mendelian disease was obtained from the OMIM database. Essential genes were defined as human–mouse orthologs associated with "abnormal survival" (except extended life span) or "sterility" in the publicly available mouse knockout data from the Mouse Genome Informatics database (MGI) (Bult et al. 2013). The list of genes associated with Complex disease was obtained from the archive of human genetic studies of complex diseases and disorders (Genetic Association Database, GAD, April 19, 2014) (Becker et al. 2004). We filtered out genes that were assigned to multiple classes.

For testing association between selection pressure and mode of inheritance, we classified exons into "autosomal recessive" and "autosomal dominant" using ClinVar (June 4, 2014) (Landrum et al. 2014). We used clinically significant mutations whose significance has been reported as "likely pathogenic" and "pathogenic" flags. Additionally, we used the combined annotation-dependent depletion scores (scaled *C*-score) to polarize deleteriousness of SNVs (Kircher et al. 2014). Specifically, we used scaled *C*-scores that rank in order of magnitude of deleteriousness all ~8.6 billion SNVs possible based on the human reference genome (hg19). For example, scaled *C*-scores of 10, 20, and 30 represent the top 10%, 1%, and 0.1% of reference genome single nucleotide variants, respectively.

Software availability

Python source code and command lines of SFS_CODE can be found in the Supplemental Material.

Acknowledgments

We thank members of the Akey laboratory for helpful discussions and feedback related to this work. J.M.A. is supported by National Institutes of Health (NIH) grants R01GM110068 and U01HG007591.

References

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56–65.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7:** 248–249.

Akey JM. 2009. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res* **19:** 711–722.

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43:** D789–D798.

Barabási A-L, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5:** 101–113.

Becker KG, Barnes KC, Bright TJ, Wang SA. 2004. The genetic association database. *Nat Genet* **36:** 431–432.

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

Bult CJ, Eppig JT, Blake JA, Kadin JA, Richardson JE, Mouse Genome Database Group. 2013. The mouse genome database: genotypes, phenotypes, and models of human disease. *Nucleic Acids Res* **41:** D885–D891.

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

Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, et al. 2014. The Reactome pathway knowledgebase. *Nucleic Acids Res* **42:** D472–D477.

Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, Merker JD, Goldfeder RL, Enns GM, David SP, et al. 2014. Clinical interpretation and implications of whole-genome sequencing. *JAMA* **311:** 1035–1045.
Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, et al. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41:** D808–D815.
Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493:** 216–220.
Fu W, Gittelman RM, Bamshad MJ, Akey JM. 2014. Characteristics of neutral and deleterious protein-coding variation among individuals and populations. *Am J Hum Genet* **95:** 421–436.
Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet* **25:** 1–5.
Gazave E, Chang D, Clark AG, Keinan A. 2013. Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. *Genetics* **195:** 969–978.
The Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res* **43:** D1049–D1056.
Glemin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015. Quantification of GC-biased gene conversion in the human genome. *Genome Res* **25:** 1215–1228.
Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* **185:** 862–864.
Gulko B, Hubisz MJ, Gronau I, Siepel A. 2015. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet* **47:** 276–283.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5:** e1000695.
Henn BM, Botigue LR, Bustamante CD, Clark AG, Gravel S. 2015. Estimating the mutation load in human genomes. *Nat Rev Genet* **16:** 333–343.
Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* **24:** 2786–2787.
Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42:** D199–D205.
Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336:** 740–743.
Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46:** 310–315.
Kotlar D, Lavner Y. 2006. The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids. *BMC Genomics* **7:** 67.
Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4:** 1073–1081.
Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42:** D980–D985.
Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, et al. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451:** 994–997.
Mathieson I, McVean G. 2012. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* **44:** 243–246.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351:** 652–654.
Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, et al. 2009. Darwinian and demographic forces affecting human protein coding genes. *Genome Res* **19:** 838–849.
Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al. 2008. Genes mirror geography within Europe. *Nature* **456:** 98–101.
Novoa EM, Pavon-Eternod M, Pan T, Ribas de Pouplana L. 2012. A role for tRNA modifications in genome structure and codon usage. *Cell* **149:** 202–213.
Racimo F, Schraiber JG. 2014. Approximation to the distribution of fitness effects across functional categories in human segregating polymorphisms. *PLoS Genet* **10:** e1004697.
Shen JC, Rideout WM III, Jones PA. 1992. High frequency mutagenesis by a DNA methyltransferase. *Cell* **71:** 1073–1080.
Siepel A, Pollard K, Haussler D. 2006. New methods for detecting lineage-specific selection. In *RECOMB'06 Proceedings of the 10th annual international conference on research in computational molecular biology*, pp. 190–205. Springer-Verlag, Berlin, Germany.
Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27:** 431–432.
Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, et al. 2005. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122:** 957–968.
Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, Raubitschek A, Ziegler S, LeProust EM, Akey JM, et al. 2013. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* **342:** 1367–1372.
Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105:** 437–460.
Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337:** 64–69.
Watterson G. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7:** 256–276.
Williamson S, Fledel-Alon A, Bustamante CD. 2004. Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics* **168:** 463–475.
Xing K, He X. 2015. Reassessing the "duon" hypothesis of protein evolution. *Mol Biol Evol* **32:** 1056–1062.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15:** 568–573.
Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol* **25:** 568–579.

Received December 7, 2015; accepted in revised form April 14, 2016.