Characteristics of Neutral and Deleterious Protein-Coding Variation among Individuals and Populations

Wenqing Fu,[1,*] Rachel M. Gittelman,[1] Michael J. Bamshad,[1,2] and Joshua M. Akey[1,*]

Whole-genome and exome data sets continue to be produced at a frenetic pace, resulting in massively large catalogs of human genomic variation. However, a clear picture of the characteristics and patterns of neutral and deleterious variation within and between populations has yet to emerge, given that recent large-scale sequencing studies have often emphasized different aspects of the data and sometimes appear to have conflicting conclusions. Here, we comprehensively studied characteristics of protein-coding variation in high-coverage exome sequence data from 6,515 European American (EA) and African American (AA) individuals. We developed an unbiased approach to identify putatively deleterious variants and investigated patterns of neutral and deleterious single-nucleotide variants and alleles between individuals and populations. We show that there are substantial differences in the composition of genotypes between EA and AA populations and that small but statistically significant differences exist in the average number of deleterious alleles carried by EA and AA individuals. Furthermore, we performed extensive simulations to delineate the temporal dynamics of deleterious alleles for a broad range of demographic models and use these data to inform the interpretation of empirical patterns of deleterious variation. Finally, we illustrate that the effects of demographic perturbations, such as bottlenecks and expansions, often manifest in opposing patterns of neutral and deleterious variation depending on whether the focus is on populations or individuals. Our results clarify seemingly disparate empirical characteristics of protein-coding variation and provide substantial insights into how natural selection and demographic history have patterned neutral and deleterious variation within and between populations.

Introduction

Mutations impose a substantial burden on fitness, disease, and longevity through the introduction of deleterious alleles into the population.[1–5] A deeper understanding of deleterious variation in humans will have profound implications for disease-mapping studies, personal genomics, and predictive medicine. A considerable amount of theoretical work has been done to inform the dynamics of deleterious variation across a range of demographic models.[6–8] Moreover, a large number of empirical studies in humans have been performed to survey patterns of deleterious variation within and between populations.[9–13] For example, in a study of 15 African American (AA) and 20 European American (EA) individuals, Lohmueller et al.[10] found that the European sample had an excess of putatively deleterious variants and through simulations demonstrated that this was most likely a consequence of the Out-of-Africa bottleneck. The proportional increase in deleterious variation in European versus African populations has also been observed in other studies.[8,14–16] Furthermore, Casals et al.[17] showed that recent founder effects in the French Canadian Quebec population have led to different patterns of deleterious variation between it and the French Canadian population. Moreover, in a clever study, Szpiech et al.[18] found that deleterious alleles were enriched in runs of homozygosity and that variable levels of inbreeding can influence patterns of deleterious variation across populations.

However, not all studies have found a clear relationship between demographic history and empirical patterns of deleterious variation. For example, Tennessen et al.[12] noted that characteristics of deleterious variants in EA and AA individuals are sensitive to how deleterious sites are defined. In addition, through detailed simulations and analyses of derived allele frequency (DAF) in a large exome sequencing data set, Simons et al.[11] suggested that the deleterious-mutation load is insensitive to recent population history and that the average number of derived alleles per individual at putatively deleterious sites is not significantly different across populations. Similarly, Do et al.[19] have recently argued that there are no differences in the per-genome accumulation of deleterious alleles across diverse human populations, which appears to contradict previous claims of differences in the proportion of deleterious variants across populations.[10,14,15]

Thus, despite the substantial amount of work that has been devoted to documenting and interpreting patterns of neutral and deleterious protein-coding variation in humans, a number of outstanding questions remain. Here, we describe a comprehensive analysis of protein-coding variation in a previously described high-coverage exome sequence data set consisting of 6,515 individuals of European and African ancestry and generated as part of the NHLBI Exome Sequencing Project (ESP).[14] Furthermore, we performed extensive simulations of neutral and deleterious variation to help interpret empirical patterns of protein-coding variation. We show that many seemingly

[1] Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [2] Department of Pediatrics, University of Washington, Seattle, WA 98195, USA
*Correspondence: wqfu@uw.edu (W.F.), akeyj@uw.edu (J.M.A.)

http://dx.doi.org/10.1016/j.ajhg.2014.09.006.

disparate observations of neutral and deleterious variation can be accounted for by opposing variation patterns that manifest depending on how variation is summarized and whether the focus is on individuals or populations. Our empirical and simulation results provide insight into how natural selection and demographic history have interacted to influence neutral and deleterious variation within and between populations.

Material and Methods

Analysis of Empirical Data

Analysis of Samples and Exome Sequencing Data

We analyzed the exomes of 6,515 individuals, including 4,298 EA individuals (1,879 males and 2,419 females) and 2,217 AA individuals (582 males and 1,635 females), from the NHLBI ESP.[14] Exome data were subjected to standard quality-control filters as previously described.[14] We further removed sites whose ancestral inference was inferred with low confidence in the six primate EPO (Enredo, Pecan, Ortheus) alignments.[20] The final data set consisted of 1,110,148 single-nucleotide variants (SNVs) in autosomes and X chromosomes.

To avoid biases caused by different sample sizes, for all population-level analyses, such as estimating the site-frequency spectrum (SFS), we randomly sampled 2,217 EA individuals to match the sample size in AA individuals. For all individual-level analyses, we defined SNVs in individuals as sites that are heterozygous or homozygous for the derived allele. We compared the per-individual number of SNVs, heterozygotes, derived homozygotes, and derived alleles between EA and AA individuals with Mann-Whitney tests. Furthermore, to account for heterogeneity in missing data among individuals, we normalized the per-individual number of derived alleles by the per-individual number of total alleles that passed filtering criteria (see Fu et al.[14]).

We also used an alternative method to account for misidentification of ancestral states. Specifically, we identified the putative ancestral state of each SNV by comparing it to the chimpanzee genome (panTro2), and we corrected ancestral misidentification by using a context-dependent mutation model.[21] In brief, this method accounts for the probability of misidentifying the ancestral state of a SNV by modeling the observed number of derived alleles (or derived homozygotes) as a mixture of SNVs whose ancestral states were correctly identified and those that were misidentified under the context-dependent substitution process.[22] In total, 1,148,406 SNVs were used in these analyses.

Moreover, we used Fisher's exact test to compare the average number of derived alleles per individual as a function of allele frequency between deleterious variants and neutral variants within populations, as well as between EA and AA populations for deleterious variants. For example, in the comparison of the enrichment of deleterious rare variants ($DAF < 0.05\%$) in one population, the elements of the 2×2 table consisted of the average per-individual number of derived alleles of rare variants ($DAF < 0.05\%$) and of variants with other frequency ($DAF \geq 0.05\%$) for both the deleterious and neutral sites.

Definition of Deleterious Variants

Quantifying evolutionary constraint through sequence conservation is widely used for identifying genomic regions that have been subject to purifying selection.[23,24] We used PhyloP scores[25] to identify putatively deleterious variants. PhyloP scores were calculated from 36 eutherian-mammal EPO alignments downloaded from the Ensembl Genome Browser (release 70) in enhanced metafile format (emf). These emf alignments were converted to multiple alignment format (maf) with the script "emf2maf.pl," also downloaded from Ensembl. Alignment blocks in maf were then sorted with the mafTools package. Finally, sorted maf alignments were converted to SS format with the msa_view program in the PHAST package. To calculate scores, we ran PhyloP (PHAST package) with the following command line option: –msa-format SS –wig-scores –mode CONACC –method LRT.

The calculation of PhyloP scores also requires a neutral model of evolution. For this, we used the phylogenetic tree provided with the 36 mammalian alignments and the substitution-rate matrix and nucleotide frequencies from the placentalMammals.mod file downloaded from the UCSC Genome Browser. PhyloP scores in wiggle (wig) format were converted to bed files with the BEDOPS package.[26] PhyloP scores were calculated with and without the human reference sequence (denoted as PhyloP_H and PhyloP_NH, respectively). Conditional on the 36-way eutherian-mammal phylogeny, simulations were performed with the base_evolve program in the PHAST package.

Population-Genetics Simulations

Forward Population Simulation for Protein-Coding Sequences

We performed forward population simulations with the program SFS_CODE[27] under different demographic models and selective regimes. We considered three general demographic models, including a population bottleneck, recent accelerated growth, and a more complicated model, by using previously inferred parameters in the EA and AA samples.[12] For the bottleneck model, a bottlenecked population was simulated from a constant population with effective size $N_e = 10{,}000$. This population experienced a bottleneck 50 ka ago, where the population size was reduced to 10% (a close approximation of the Out-of-Africa bottleneck)[28] and 1% of the original size, and recovered from the bottleneck 25 ka ago. The Out-of-Africa bottleneck has also been modeled as a shorter and more severe bottleneck.[29] In this model, a constant population ($N_e = 10{,}000$) experienced a bottleneck 118 ka ago and a quick recovery 108 ka ago. During the bottleneck, the population size was reduced to 7.57% of the original. We also simulated data under this model to study how robust our results are to particular implementations of the Out-of-Africa bottleneck.

For the model of recent population growth, a population started expanding from a constant population ($N_e = 10{,}000$) 5 ka ago. We considered different growth rates, including 0%, 2.0% (a close approximation for the population with European ancestry),[12] and 3.0% per generation.

In the more realistic demographic model, European and African populations split 51 ka ago, and the European lineage incurred two bottleneck events (the Out-of-Africa bottleneck 51 ka ago and the split of non-African populations 23 ka ago) and an initial population expansion with a growth rate of 0.307% per generation, whereas the African population evolved as a constant population during this period.[28] Beginning 5.115 ka ago, accelerated population growth occurred for both European and African populations with growth rates of 1.95% and 1.66%, respectively.[12] The simulated AA population is a result of recent admixture from European (20%) and African (80%) populations.

A total of 2,500 individuals were simulated for each parameter combination. For each individual, we simulated 5,000 independent genes, each with four 500 bp exons that are equally spaced with 2,000 bp introns (sequences for the introns were not