Neutral Theory, Disease Mutations, and Personal Exomes

Sudhir Kumar*[,1,2,3] and Ravi Patel[1,2]

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA

[2]Department of Biology, Temple University, Philadelphia, PA

[3]Center for Excellence in Genome Medicine and Research, King Abdulaziz University, Jeddah, Saudi Arabia

*Corresponding author: E-mail: s.kumar@temple.edu.

Associate editor: Heather Rowe

Abstract

Genetic differences between species and within populations are two sides of the same coin under the neutral theory of molecular evolution. This theory posits that a vast majority of evolutionary substitutions, which appear as differences between species, are (nearly) neutral, that is, these substitutions are permitted without a significantly adverse impact on a species' survival. We refer to them as evolutionarily permissible (ePerm) variation. Evolutionary permissibility of any possible variant can be inferred from multispecies sequence alignments by applying sophisticated statistical methods to the evolutionary tree of species. Here, we explore the evolutionary permissibility of amino acid variants associated with genetic diseases and those observed in personal exomes. Consistent with the predictions of the neutral theory, disease associated amino acid variants are rarely ePerm, much more biochemically radical, and found predominantly at more conserved positions than their non-disease counterparts. Only 10% of amino acid mutations are ePerm, but these variants rise to become two-thirds of all substitutions in the human lineage (a 6-fold enrichment). In contrast, only a minority of the variants in a personal exome are ePerm, a seemingly counterintuitive pattern that results from a combination of mutational and evolutionary processes that are, in fact, broadly consistent with the neutral theory. Evolutionarily forbidden variants outnumber detrimental variants in individual exomes and may play an underappreciated role in protecting against disease. We discuss these observations and conclude that the long-term evolutionary history of species can illuminate functional biomedical properties of variation present in personal exomes.

Key words: **phylomedicine, diseases, long-term history.**

Introduction

The proposal of the neutral theory of molecular evolution in the late 1960s (Kimura 1968; King and Jukes 1969) was a culmination of many years of observations of amino acid sequence differences among species and protein polymorphisms within populations. The neutral theory weaves patterns of within and among species differences into the same evolutionary framework in which negative selection and random genetic drift are the central processes. This proposal was in stark contrast to then prevalent theories that embraced positive selection as the primary force driving evolution (reviewed in Nei et al. 2010). Over the last five decades, the neutral theory has successfully explained much of the observed genetic variation at various levels of biological organization, including populations, species, genes, genomes, and gene regulation (Lynch 2007; Ho et al. 2017). It has also served as the fundamental basis for the development of methods to discover signatures of positive selection at the molecular level (Nielsen 2005; Booker et al. 2017). It has been argued that the neutral theory is a null hypothesis for genomic variation associated with diseases and complex traits (Dudley et al. 2012; Williams et al. 2016), and measures of evolutionary conservation are frequently used to assess deleteriousness of novel mutations in phylomedicine (Ramensky et al. 2002; Ng and Henikoff 2003; Kumar et al. 2011, 2012).

Many excellent reviews about the history and the current state of the neutral theory have been published, so here we focus on genetic variation associated with human diseases and personal exomes. Disease mutations represent negative selection in action and personal genomes are the ultimate subjects of selection. In this perspective, we will only consider protein sequence variation, specifically single nucleotide variants resulting in missense mutations, for two reasons. First, Motoo Kimura's early work was based on comparative studies of the amino acid sequences of homologous proteins among related organisms (Kimura 1983: 306). Second, the molecular bases of disease caused by amino acid variation is much better understood than disease associated with noncoding variants in which the translation from correlation to causation remains a major challenge. In the following, we refer to amino acid variants associated with human diseases as disease variants for brevity.

Fixation of (Selectively) Neutral Alleles (Molecular Evolution)

That the random fixation of selectively neutral mutants causes the great majority of evolutionary substitutions is the primary tenet of the neutral theory (Kimura 1983: 306). This tenet predicts that evolutionary fitness will not be significantly impacted by a mutation at a position if that mutant

Perspective

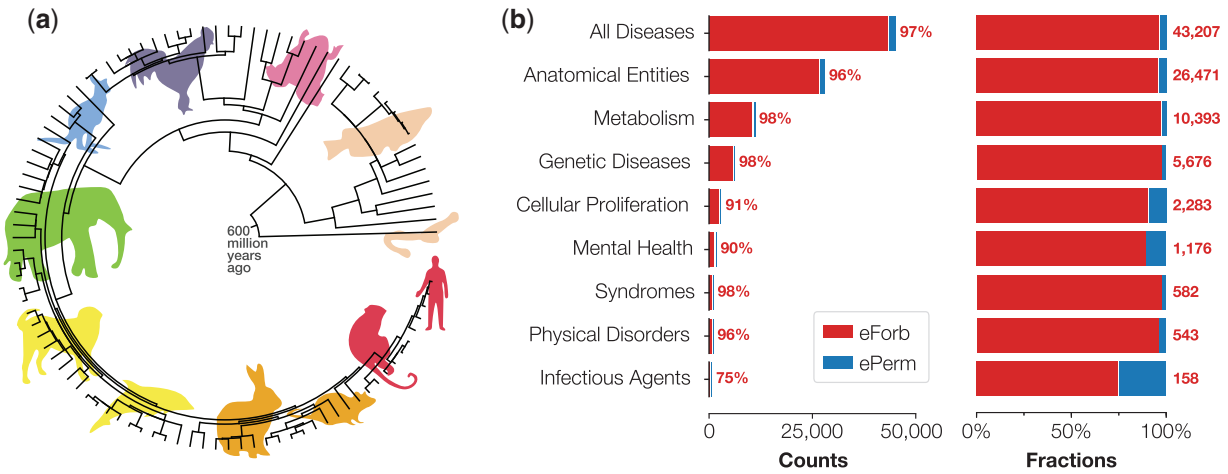

Fig. 1. Between-species differences predict within-species variation. (*a*) The TimeTree of 100 species used in calculating the evolutionary probabilities following Liu et al. (2016). (*b*) The absolute counts and relative fractions of missense disease variants that are evolutionarily permissible (ePerm; blue) and evolutionarily forbidden (eForb; red) are shown. Evolutionary probabilities (EPs) were calculated using orthologous amino acid sequence alignments for 100 vertebrate species for 18,835 proteins downloaded from the UCSC browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/multiz100way/alignments; last accessed July 12, 2016). Species divergence times were retrieved from the TimeTree database (Kumar et al. 2017). A total of 44,685 disease associated missense variants were retrieved from HGMD (Stenson et al. 2009); a variant was designated eForb if it had an EP < 0.05 and ePerm otherwise. Disease categories are top-level Disease Ontology terms (Kibbe et al. 2015). Phenotypes found in HGMD were linked to disease ontology terms using NCBI MedGen Concept Unique Identifiers. Each disease variant may be found in multiple disease ontology classes. Because the number of variants analyzed was large in every disease category, the difference in the fraction of variants that are eForbs was significant for all disease categories compared with other categories ($P < 10^{-5}$), but the absolute difference is small in many cases, except for diseases by infectious agents.

allele is observed at a homologous position in other species, as long as the function of that position has not changed. That is, the mutant allele will be evolutionarily permissible (ePerm). All amino acid residues not observed among species at a position are then evolutionarily forbidden (eForb), which are expected to be lethal or decrease reproductive fitness by appearing as disease mutations. Indeed, <20% of the disease variants are ePerm, that is, they do not appear in orthologous positions in any of the 99 other vertebrate species examined; see also Miller and Kumar (2001); Kondrashov et al. (2002).

A more sophisticated approach to quantifying the evolutionary permissibility of a mutant allele is to compute the Bayesian evolutionary probability (EP) of an amino acid allele by using a multispecies sequence alignment and the species phylogeny (Liu et al. 2016) (fig. 1a). Variants with a low EP would be eForb, as such variants have been eliminated over evolutionary time by negative selection under the neutral theory framework. A variant will receive a high EP if it is ePerm, that is, observed among species. Importantly, EP of a variant at a position is not influenced by the human alleles found at that position (fixed or polymorphic). It is preferable to use the EP approach to categorize variant alleles into eForbs and ePerms, rather than a simple alignment scanning approach, because EP is better capable of assessing disease variants that have arisen independently in species distantly related to humans (Liu et al. 2016). In all further analysis and discussion, we refer to variants with EP ≥ 0.05 to be ePerm and variants with EP < 0.05 to be eForb. Patel et al. (2018) showed that alleles with EP < 0.05 are decidedly unexpected in humans under neutral evolution.

Greater than 97% of the disease variants are eForbs (fig. 1*b*), which is consistent with the role of purifying selection in the neutral theory. All broad disease categories are associated with a similar fraction of eForbs, except that only 75% of the variants for infectious diseases are eForbs. This is likely because the diseases by infectious agents are subject to pressures from host–pathogen interactions that are simultaneously driven by both human (host) and external (pathogen) genetic factors, introducing a co-evolutionary dynamic not present in human-driven diseases (those primarily attributed to human mutation factors). ePerm variants are more rarely involved in pathogenesis for diseases stemming from fundamental biological organization levels (molecular and cellular) as compared with more derived levels of biological structures, including tissues and organs; for example, diseases of metabolism are associated with significantly fewer ePerms than are diseases of anatomical entities (2% vs. 4%, respectively; *z*-test $P \ll 10^{-5}$).

The neutral theory also predicts that evolutionary substitutions that are not disruptive to the existing structure and function of a molecule will occur more readily (Kimura 1983: 312). Through this prediction, Kimura connected evolutionary constraints with the structure and function of proteins. One can measure functional disruptiveness of an amino acid mutation by using Grantham distance, which quantifies differences in volume, composition, and polarity between amino acids (Grantham 1974). The higher the Grantham distance between the original and mutant amino acid, the greater the chance of functional impact. Amino acid substitutions among species show the lowest Grantham distance for amino acid positions that are the most conserved (fig. 2*a*), which

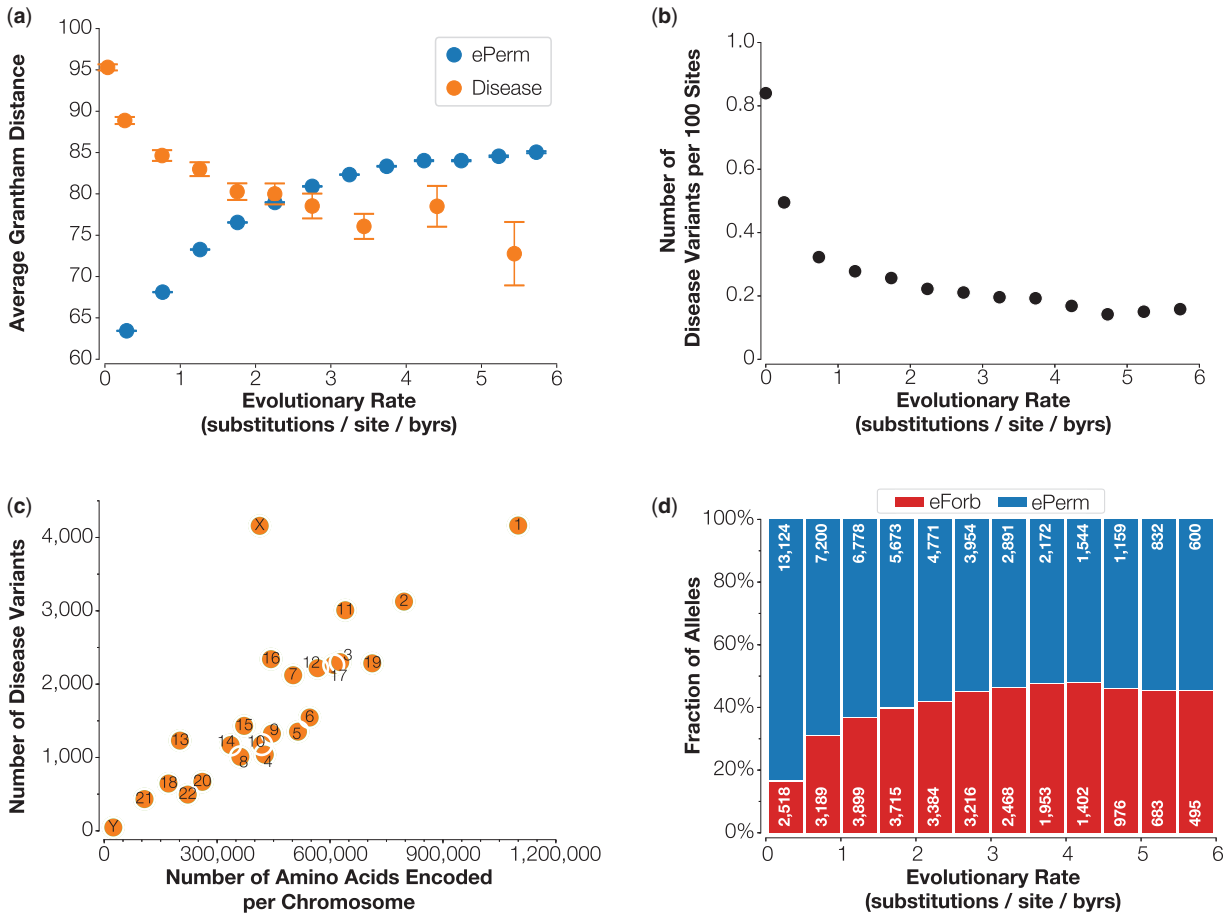

Fig. 2. Evolutionary properties of disease associated missense variants. (*a*) Biochemical severity (Grantham distance) of evolutionarily permissible (ePerm) substitutions and disease associated variants at positions with different rates of evolution. Error bars show the standard error of the mean. (*b*) Number of disease variants found per 100 sites evolving with different evolutionary rates, from highly conserved (low rates) to highly variable (high rates). (*c*) The relationship of the number of disease variants with the number of amino acid sites encoded by each human chromosome. Circles are labeled with the corresponding chromosome number. (*d*) The proportions of evolutionarily permissible (ePerm) and forbidden (eForb) alleles that have been fixed in the human exome since the human–chimpanzee divergence. Results from different evolutionary rate categories are shown. Long-term evolutionary rates were calculated following the procedure in Kumar et al. (2009). Evolutionary rates are binned in increments of 0.5, with values for all constant sites plotted at rate = 0. Average Grantham distance for ePerms is calculated over all substitutions (with respect to the human protein) that are observed in orthologous positions in the other nonhuman vertebrates. Grantham distance for disease variants is calculated based on the wild type to disease-associated mutant change.

means that the number of substitutions as well as their biochemical disruptiveness is lower at positions that have experienced stronger constraint over evolutionary time. The timing of the onset of disease would also modulate the evolutionary neutrality of amino acid variants, because early onset diseases have a greater potential to impact fitness than late onset diseases that will likely have small effects on fecundity and reproductive success. As expected, variants associated with early onset diseases show up to 17% greater biochemical disruptiveness than those associated with late onset diseases (Subramanian and Kumar 2006).

It naturally follows that amino acid positions under stronger evolutionary constraints will be more likely to be implicated in disease (Kimura 1983: 306). Indeed, the incidence of disease mutations is the highest in the most highly conserved positions (fig. 2*b*; Miller and Kumar 2001; Kumar et al. 2009). Also, the X-chromosome shows a much greater incidence of disease variants (fig. 2*c*), because unlike autosomes, X chromosomes are effectively present in only one copy (male hemizygosity, female X-inactivation; Kimura 1983: 104; Li et al. 2010). In these hemizygous genotypes, genes harboring recessive detrimental variants will be subject to greater purifying selection than autosomes.

Despite contributing only a small portion of known disease variation, ePerms sometimes are detected as human disease variants. This can be for many different reasons. First, and foremost, we expect some currently annotated disease mutations to be linked to the functional variant, rather than being causal themselves. Second, compensatory evolution may have altered the effect of the variants (Kondrashov et al. 2002; Subramanian and Kumar 2006; Xu and Zhang 2014). Third, previously neutral alleles may have become detrimental in humans due to genomic (e.g., epistasis) or external environmental changes, a possibility that was considered by Kimura (1983). Finally, as previously mentioned, some diseases may not affect reproductive success. These possibilities

notwithstanding, an overwhelming majority of human disease mutations is predicted to be deleterious based on evolutionary patterns observed among species, which is consistent with the importance of negative selection advocated by the neutral theory.

Fixation of Human-Specific Mutations (Recent Molecular Evolution)

A comparison of the human exome with apes and other primates reveals that >85,000 amino acid mutations have become fixed in the human lineage, after the human–chimpanzee species divergence. A majority of these human-specific evolutionary substitutions is ePerms (~65%), but the actual count of eForbs is large (~30,000). Small ancestral human population size and complex demographic histories likely led to the fixation of many weakly deleterious mutations, a conjecture supported by the population genetic analysis of modern humans (Henn et al. 2015). We found that these eForbs are moderately radical biochemically, as their average Grantham distance (~82) is slightly higher than that for ePerms (~69), but lower than that for disease variants (~91) (Miller and Kumar 2001). As expected, the lowest proportion of eForbs are found at completely conserved positions (~15%, fig. 2*d*), indicating that less conserved sites are more likely to harbor weakly deleterious variants.

Many eForb substitutions may also be involved in adaptation, because eForbs may actually change the functional properties of the protein, rather than simply rendering them inactive. Liu et al. (2016) report ten positions where a reversion to an ePerm allele is associated with disease. That is, the (unknown) functional change enabled by the eForb was beneficial or compensatory. Patel et al. (2018) also report >250 eForbs that occur with high frequency in human populations at codons where reversion to an ePerm is associated with a detrimental phenotype. Kimura (1983: 115) noted that the fixation of weakly deleterious mutations by random drift is essentially equivalent to a change in environment, which would require occasional adaptive substitutions to avoid species extinction. However, the discovery of adaptive variants needed to counteract the detrimental effects of so many weakly deleterious polymorphisms and fixations remains challenging.

Selective Neutrality of Intraspecific Variation (Population Genetics)

Kimura (1983: 299) indicated that the analysis of rare variant alleles is important for understanding the mechanisms that maintain genetic variability. Testaments to the importance of this foresight are the growing visibility of rare variants in complex diseases and traits (Cirulli and Goldstein 2010) and the critical insight into fundamental mutational patterns provided by analyses of rare variants (Schaibley et al. 2013; Carlson et al. 2017). In the 1,000 Genomes Project Phase 3 (1KG) data (1000 Genomes Project Consortium et al. 2015), a vast majority of identified variants is singletons (fig. 3*a*). Singletons, by definition, are the rarest variants in the human genome, and a vast majority of these potentially new mutations is expected to be detrimental (Kryukov et al. 2007) and recessive (García-Dorado and Caballero 2000). Indeed, singletons are overwhelmingly eForbs (fig. 3*b*); only 8% of them are ePerm (23,822 out of 301,523 variants). This observation is consistent with an early suggestion that ~9% of amino acid altering mutations are selectively neutral (Kimura and Ohta 1973). Variants that occur twice in a population (doubletons) are also overwhelmingly eForbs, but their distribution is shifted slightly towards higher EPs (fig. 3*b*). This means that purifying selection has already eliminated singleton eForbs before they increased in frequency. Therefore, eForbs are subject to natural selection over time, and fewer and fewer eForb alleles increase in frequency. This results in a greater fraction of higher frequency variants identified as ePerm (fig. 3*c*; Liu et al. 2016).

More than 60,000 distinct variants show minor allele frequencies >1% in at least one of the five major continental supergroups defined in the 1KG data. Of these, <100 missense variants have been implicated in adaptive evolution (Dudley et al. 2012; Patel et al. 2018), which is consistent with the minor role that neutral theory posits for positive selection. However, it has been difficult to identify adaptation at the molecular level, because many phenotypic adaptations are likely to involve multiple genes, each of which may transmit weak signals of selection (Pritchard et al. 2010; Hernandez et al. 2011). If this is so, many truly beneficial variants will be selectively neutral from the neutral theory perspective and their individual fates will be determined by genetic drift (Kimura 1983: 296).

Some eForbs may have become neutral or adaptive in humans due to an altered environment, possibly coupled with the evolution of new epistatic interactions. Kimura already considered that mutant alleles at different loci may be individually harmful, but neutral or adaptive in combination (Kimura and Maruyama 1966). However, such events may be more frequent than Kimura's expectations. Also, high frequency eForbs may have been adaptive in early human history, but became detrimental due to a change in environment. This class of variants will be consistent with the mismatch theory (Nesse and Williams 1994), which suggests that complex diseases are caused by previous adaptations that have become detrimental in contemporary human environments following the agricultural and technological revolutions ("maladaptations"). This hypothesis is testable at the molecular level, because genetic factors would be required to propagate adaptive signatures through generations. However, the genetic underpinnings of biological adaptations are notoriously difficult to unravel, with a few classic exceptions (Jablonski and Chaplin 2000; Bersaglieri et al. 2004). In the future, the relative contributions of new mutations, previously adaptive variants, and previously neutral variants to disease phenotypes will improve our assessment of the mismatch theory at the molecular level.

From Species and Populations to Individuals (Personal Exomes)

A human exome contains over 10 million amino acid positions, which is the summation of amino acid sequence

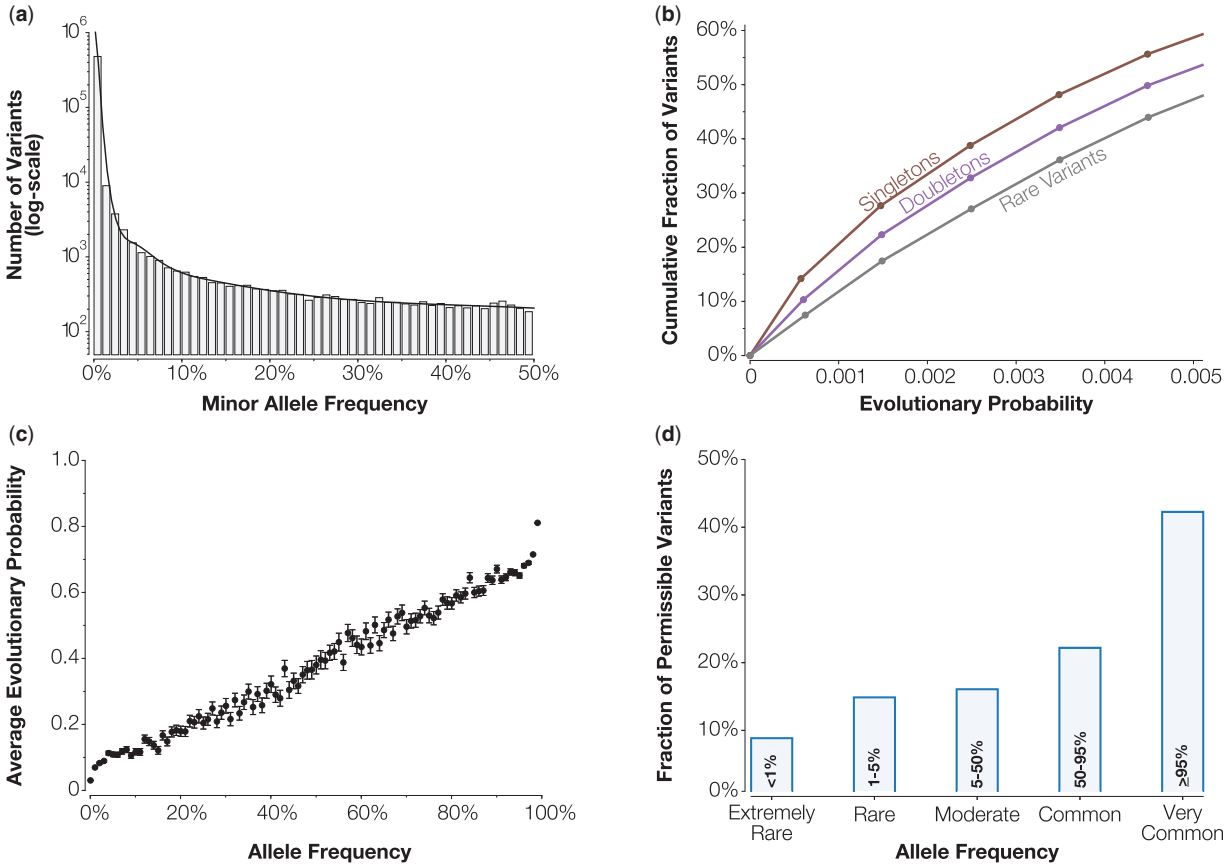

FIG. 3. Population features of human missense variation. (*a*) Distribution of minor alleles frequencies in the 1000 Genomes Project Phase 3 (1KG) data (1000 Genomes Project Consortium et al. 2015). (*b*) Cumulative distribution of evolutionary probabilities (EP) for variants found in one (singletons), two (doubletons), and a few (3–49) copies in the 1KG sample. Each fraction is plotted in the middle of the corresponding EP bin of width 0.001. See figure 1 for details on EP calculation. (*c*) Relationship between EP and population allele frequency (AF) for all missense variants found in the 1KG data. Variants are binned by AF in 1% increments, and plotted against average EP. Error bars show the standard error of the mean. (*d*) The fraction of evolutionarily permissible variants found in various AF classes: extremely rare (<1%), rare (1–5%), moderate (5–50%), common (50–95%), and very common (>95%).

lengths of all proteins encoded. In the 1KG data and 100 vertebrate species, we estimated that ~100,000 (1%) positions in an average personal exome harbor an amino acid variant that arose after the human–chimpanzee divergence (human-derived). Of these, ~86,000 are evolutionary substitutions that are found as fixed differences in modern humans, of which two-thirds are ePerms (fig. 4*a*). Interestingly, less than a quarter of identified polymorphisms are ePerms (1,862 out of 9,085; fig. 4*a*). That is, an overwhelming majority of variants in a personal exome is eForbs. This seemingly counterintuitive pattern is actually consistent with the neutral theory and can be explained by a simple example. Consider a population of 100 individuals that has 102 polymorphic amino acid positions: two harboring eForbs (adaptive or deleterious) that have arisen to 50% allele frequency due to selection or demography, and 100 positions harboring ePerms that occur with a frequency of 1% each. At the population level, we would conclude that there are 50 times more ePerms than eForbs, which would be consistent with neutral expectations. However, on average, a personal exome will contain only one eForb and one ePerm, so the ratio in a personal exome will be 1:1, rather than 50:1. Therefore, a stark difference exists between the personal and evolutionary perspectives on genome variation.

Through the lens of evolutionary history, we explored the relative fractions of homozygous and heterozygous variants that are eForb in an average personal exome. 76% of the homozygotes were eForbs (3,077 out of 4,040), which is only slightly lower than that for heterozygotes (82%; 4,416 out of 5,045; fig. 4*a*). 55 homozygous and 187 heterozygous eForbs match a disease variant in the Human Gene Mutation Database (HGMD; Stenson et al. 2009), which means that a large number of known disease variants exist in our personal exomes, as has been previously reported (Tennessen et al. 2012). Interestingly, 120 (out of 3,077) homozygous eForbs appear to be protective, because the *absence* of the eForb allele is associated with a disease in the HGMD data set (GWAS odds ratio < 1). These homozygous protective eForbs (120) outnumber homozygous detrimental variants (55), in which the *presence* of eForb alleles is associated with disease (GWAS odds ratio > 1; fig. 4*b*). But, heterozygotes show an opposite pattern in which detrimental variants are twice the number of protective variants (fig. 4*b*). If we assume that the patterns observed using this limited data set

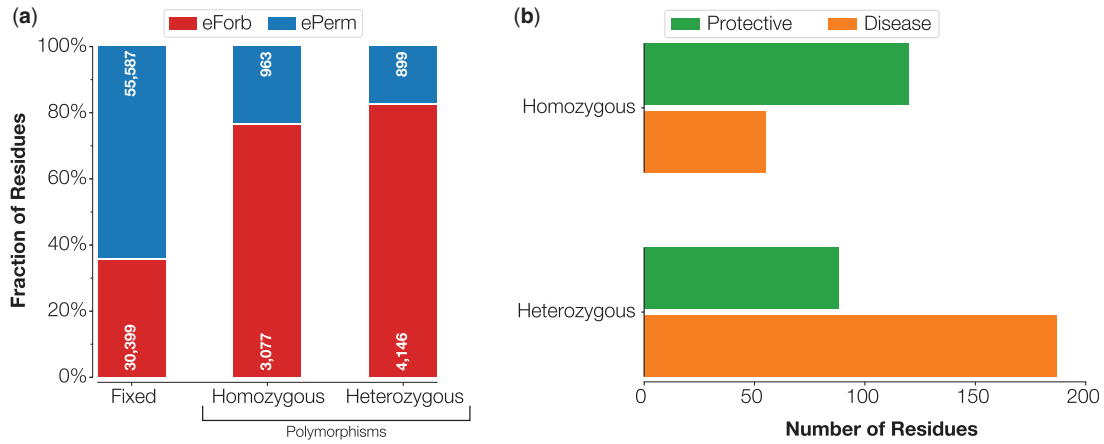

FIG. 4. Evolutionary permissibility of missense variants in an average personal exome. (*a*) Proportions of evolutionarily permissible (ePerm; blue) and evolutionarily forbidden (eForb; red) residues found at fixed and polymorphic sites. Because of the large number of variants in an exome, the small 6% difference in eForbs as homozygotes compared with heterozygotes is statistically significant (*z*-test $P \ll 10^{-5}$). (*b*) Number of disease and protective eForb variants found at homozygous and heterozygous positions. ePerm and eForb residue counts were averaged across each of the 2,504 individuals found in the 1000 Genomes Phase 3 (1KG) data (1000 Genomes Project Consortium et al. 2015). A variant was designated as "protective" if the lack of that allele in the "cases" from case–control studies was associated with disease (i.e., GWAS odds ratio < 1.0), otherwise the allele was designated to be a disease variant. All alleles that arose in the human lineage, after the human–chimpanzee divergence, were considered derived alleles (i.e., derived alleles were not observed in other great ape species).

are an indication of relative numbers of homozygous protective and detrimental variants in the whole exome, we would conclude that the number of protective variants exceeds the number of detrimental variants in an exome. This could explain why our exomes can carry so many disease associated and eForb variants.

We must acknowledge the limited information available to support genotype–phenotype associations, as we were able to find GWAS odds ratios for only 175 out of 3,077 homozygous eForbs in the HGMD data set. Because of the polygenic nature of most complex phenotypes, the strength of their biological effect and, thus, our power of discovery is expected to be greatly impacted by epistasis, pleiotropy, and an individual's physical environment. Some eForbs may counteract deleterious effects in the context of another eForb (sign epistasis, Weinreich et al. 2005), have both deleterious and beneficial effects on fitness (e.g., pleiotropy, Patel et al. 2018), and even have different effects in alternate environments (phenotypic plasticity, Bradshaw 1965). Unless the variant in question is studied in the multiple contexts, it can easily be overlooked in association studies. Furthermore, the effect size may be too small to be detectable without large sample sizes. The growing number of whole exome sequences and electronic health records will help to disentangle many of these confounding processes (Green and Guyer 2011). We predict that understanding the true nature of the interactions between opposing protective and detrimental eForbs will be an important avenue for research in personalized medicine. This endeavor will be greatly aided by integrative analysis of intra and interspecific differences and by the increasing resolution of the biological interactomes.

Conclusion

In this perspective, we have used measurements of the evolutionary permissibility of variants, derived from multispecies protein alignments, to evaluate disease variants, population polymorphisms, and personal exome mutations. Our examples make it clear that the macroevolution of protein sequences, consistent with the neutral theory, is highly consistent with observed patterns of disease associated variation. This means that processes of negative selection and genetic drift are important to understanding biomedically relevant variation, contradicting Gluckman et al.'s (2016: xv) suggestion that the long-term evolutionary history captured in the macroevolution of species is not necessary for the field of evolutionary medicine.

Whereas populations and species consist of individuals, individuals are not the primary focus of theoretical discussions. This contrasts sharply with the role of an individual in medicine; a medical doctor's paramount concern is the patient they are treating. Also, the term "fitness" in molecular evolutionary theories does not have the same meaning in regards to personal health and disease where one instead refers to the wellness of an individual. In the field of genomic medicine, there is high interest in developing polygenic risk scores that integrate risks associated with the presence of deleterious variants in individuals (Euesden et al. 2015; Lewis and Vassos 2017). We hope that measuring evolutionary permissiveness will help to generate discriminative polygenic risk scores, as the evolutionary permissibility of patterns observed in personal exomes, variation accumulated in populations, and differences among species are part of a continuum in which selection and random genetic drift play a central role, as envisioned by the proponents of the neutral theory.

Acknowledgments

We thank Laura Scheinfeldt, Rob Kulathinal, Heather Rowe, and Qiqing Tao for critical comments. This work was supported by grants from the National Institutes of Health to S.K. (LM012487).

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* 526(7571):68–74.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet*. 74(6):1111–1120.

Booker TR, Jackson BC, Keightley PD. 2017. Detecting positive selection in the genome. *BMC Biol*. 15(1):98.

Bradshaw AD. 1965. Evolutionary significance of phenotypic plasticity in plants. Adv Genet. 13:115–155.

Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, Myers RM, Boehnke M, Kang HM, Scott LJ, Li JZ, et al. 2017. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. Available from: http://dx.doi.org/10.1101/108290

Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet*. 11(6):415–425.

Dudley JT, Kim Y, Liu L, Markov GJ, Gerold K, Chen R, Butte AJ, Kumar S. 2012. Human genomic disease variants: a neutral evolutionary explanation. *Genome Res*. 22(8):1383–1394.

Euesden J, Lewis CM, O'Reilly PF. 2015. PRSice: polygenic Risk Score software. *Bioinformatics* 31(9):1466–1468.

García-Dorado A, Caballero A. 2000. On the average coefficient of dominance of deleterious spontaneous mutations. *Genetics* 155(4):1991–2001.

Gluckman P, Beedle A, Buklijas T, Low F, Hanson M. 2016. Principles of evolutionary medicine. New York: Oxford University Press.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185(4154):862–864.

Green ED, Guyer MS, National Human Genome Research Institute. 2011. Charting a course for genomic medicine from base pairs to bedside. *Nature* 470(7333):204–213.

Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S. 2015. Estimating the mutation load in human genomes. *Nat Rev Genet*. 16(6):333–343.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, 1000 Genomes Project, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331(6019):920–924.

Ho W-C, Ohya Y, Zhang J. 2017. Testing the neutral hypothesis of phenotypic evolution. *Proc Natl Acad Sci U S A*. 114(46):12219–12224.

Jablonski NG, Chaplin G. 2000. The evolution of human skin coloration. *J Hum Evol*. 39(1):57–106.

Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D, et al. 2015. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res*. 43(D1):D1071–D1078.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217(5129):624–626.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press

Kimura M, Maruyama T. 1966. The mutational load with epistatic gene interactions in fitness. *Genetics* 54(6):1337–1351.

Kimura M, Ohta T. 1973. Mutation and evolution at the molecular level. *Genetics* 73(Suppl 73)19–35.

King JL, Jukes TH. 1969. Non-Darwinian evolution. *Science* 164(3881):788–798.

Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky–Muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A*. 99(23):14878–14883.

Kryukov GV, Pennacchio LA, Sunyaev SR. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet*. 80(4):727–739.

Kumar S, Dudley JT, Filipski A, Liu L. 2011. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends Genet*. 27(9):377–386.

Kumar S, Sanderford M, Gray VE, Ye J, Liu L. 2012. Evolutionary diagnosis method for variants in personal exomes. *Nat Methods*. 9(9):855–856.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol*. 34(7):1812–1819.

Kumar S, Suleski MP, Markov GJ, Lawrence S, Marco A, Filipski AJ. 2009. Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. *Genome Res*. 19(9):1562–1569.

Lewis CM, Vassos E. 2017. Prospects for using risk scores in polygenic medicine. *Genome Med*. 9(1):96.

Liu L, Tamura K, Sanderford M, Gray VE, Kumar S. 2016. A molecular evolutionary reference for the human variome. *Mol Biol Evol*. 33(1):245–254.

Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, Albrechtsen A, Andersen G, Cao H, Korneliussen T, et al. 2010. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet*. 42(11):969–972.

Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A*. 104(Suppl 1):8597–8604.

Miller MP, Kumar S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. *Hum Mol Genet*. 10(21):2319–2328.

Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. *Annu Rev Genomics Hum Genet*. 11:265–289.

Nesse RM, Williams GC. 1994. Why we get sick: the new science of Darwinian medicine. New York: Random House

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 31(13):3812–3814.

Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Genet*. 39:197–218.

Patel R, Sanderford MD, Lanham TR, Tamura K, Platt A, Gilksberg BS, Dudley JT, Xu K, Scheinfeldt LB, Kumar S. 2018. Adaptive landscape of protein variation in human exomes. Available from: http://dx.doi.org/10.1101/282152

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*. 20(4):R208–R215.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*. 30(17):3894–3900.

Schaibley VM, Zawistowski M, Wegmann D, Ehm MG, Nelson MR, St Jean PL, Abecasis GR, Novembre J, Zöllner S, Li JZ. 2013. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. *Genome Res*. 23(12):1974–1984.

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The human gene mutation database: 2008 update. *Genome Med*. 1(1):13.

Subramanian S, Kumar S. 2006. Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC Genomics*. 7:306.

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(6090):64–69.

Weinreich DM, Watson RA, Chao L. 2005. Perspective: sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* 59(6):1165–1174.

Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. 2016. Identification of neutral tumor evolution across cancer types. *Nat Genet*. 48(3):238–244.

Xu J, Zhang J. 2014. Why human disease-associated residues appear as the wild-type in other species: genome-scale structural evidence for the compensation hypothesis. *Mol Biol Evol*. 31(7):1787–1792.