The population genetics of human disease: the case of recessive, lethal mutations

Carlos Eduardo G. Amorim[1,2*], Ziyue Gao[3], Zachary Baker[4], José Francisco Diesel[5], Yuval B. Simons[1], Imran S. Haque[6], Joseph Pickrell[1,7+], Molly Przeworski[1,4+]

[1] Department of Biological Sciences, Columbia University, New York, NY 10027

[2] CAPES Foundation, Ministry of Education of Brazil, Brasília, DF, Brazil 70040.

[3] Howard Hughes Medical Institution, Stanford University, Stanford, CA 94305

[4] Department of Systems Biology, Columbia University, New York, NY 10027

[5] Universidade Federal de Santa Maria, Santa Maria, RS, Brazil 97105

[6] Counsyl, 180 Kimball Way, South San Francisco, CA

[7] New York Genome Center, New York, NY 10013

[+]These authors co-supervised this work

[*] To whom correspondence should be addressed (guerraamorim@gmail.com)

Abstract

Do the frequencies of disease mutations in human populations reflect a simple balance between mutation and purifying selection? What other factors shape the prevalence of disease mutations? To begin to answer these questions, we focused on one of the simplest cases: recessive mutations that alone cause lethal diseases or complete sterility. To this end, we generated a hand-curated set of 417 Mendelian mutations in 32 genes, reported to cause a recessive, lethal Mendelian disease. We then considered analytic models of mutation-selection balance in infinite and finite populations of constant sizes and simulations of purifying selection in a more realistic demographic setting, and tested how well these models fit allele frequencies estimated from 33,370 individuals of European ancestry. In doing so, we distinguished between CpG transitions, which occur at a substantially elevated rate, and other mutation types. Whereas observed frequencies for CpG transitions are close to expectation, the frequencies observed for other mutation types are an order of magnitude higher than expected; this discrepancy is even larger when subtle fitness effects in heterozygotes or lethal compound heterozygotes are taken into account. In principle, higher than expected frequencies of disease mutations could be due to widespread errors in reporting causal variants, compensation by other mutations, or balancing selection. It is unclear why these factors would affect CpG transitions differently from other mutations, however. We argue instead that the unexpectedly high frequency of non-CpGti disease mutations likely reflects an ascertainment bias: of all the mutations that cause recessive lethal diseases, those that by chance have reached higher frequencies are more likely to have been

identified in medical studies and thus to have been included in this study. Beyond the specific application, this study highlights the parameters likely to be important in shaping the frequencies of Mendelian disease alleles.

Author Summary

What determines the frequencies of disease mutations in human populations? To begin to answer this question, we focus on one of the simplest cases: mutations that cause completely recessive, Mendelian disease. We first review theory about what to expect from mutation and selection in a population of finite size and further generate predictions based on simulations using a realist demographic scenario of human evolution. For a highly mutable type of mutations, involving transitions at CpG sites, we find that the predictions fit observed frequencies of recessive lethal disease mutations well. For less mutable types, however, predictions tend to under-estimate the observed frequency. We discuss possible explanations for the discrepancy and point to a complication that, to our knowledge, is not widely appreciated: that there exists ascertainment bias in disease mutation discovery. Specifically, we suggest that alleles that have been identified to date are likely the ones that by chance have drifted to higher frequencies and are thus more likely to have been mapped. More generally, our study highlights the parameters that influence the frequencies of Mendelian disease alleles, helping to interpret the relevance of variants of unknown significance based on their allele frequencies.

Introduction

New disease mutations arise in heterozygotes and either drift to higher frequencies or are rapidly purged from the population, depending on the strength of selection and the demographic history of the population [1-6]. Elucidating the relative contributions of mutation, natural selection and genetic drift will help to understand why disease alleles persist in humans. Answers to these questions are also of practical importance, in informing how genetic variation data can be used to identify additional disease mutations [7].

In this regard, rare, Mendelian diseases, which are caused by single highly penetrant and deleterious alleles, are perhaps most amenable to investigation. A simple model for the persistence of mutations that lead to Mendelian diseases is that their frequency is determined by a balance between their introduction by mutation and elimination by purifying selection, i.e., that we expect to find them at "mutation-selection balance" (MSB) [4]. In finite populations, random drift leads to stochastic changes in the frequency of any mutation, so demographic history, in addition to mutation and natural selection, plays an important role in shaping the frequency distribution of deleterious mutations [3].

Another factor that may be important in shaping the frequency of highly penetrant disease mutations is genetic interactions. For instance, a disease mutation may be rescued by another mutation in the same gene [8-10] or by a modifier locus elsewhere in the genome that modulates the severity of the disease symptoms or the penetrance of the disease allele (e.g. [11-13]).

For a subset of disease alleles that are recessive, an alternative model for their frequency in the population is that there is an advantage to carrying one copy but a disadvantage to carrying two or none, such that the alleles persist due to overdominance, a form of balancing selection. Well known examples include sickle cell anemia, thalassemia and G6PD deficiency in populations living where the malaria is endemic [14]. Beyond these cases, the importance of overdominance in maintaining the high frequency of disease mutations is unknown.

Here, we tested hypotheses about the persistence of mutations that cause lethal, recessive, Mendelian disorders. This case provides a good starting point towards elucidating the determinants of disease mutations in human populations, because a large number of recessive lethal disorders have been mapped (e.g., genes have already been associated with >66% of Mendelian disease phenotypes; [15]) and, while the fitness effects of most diseases are hard to estimate, for lethal diseases, the selection coefficient is clearly 1 for homozygote carriers (at least in the absence of modern medical care). Moreover, the simplest expectation for mutation-selection balance would suggest that, given a per base pair mutation rate on the order of 10^{-8} per generation [16], the frequency of such alleles would be $\sqrt{u}$, i.e., $\sim 10^{-4}$ [4]. Thus, sample sizes in human genetics are now sufficiently large that we should be able to observe recessive disease alleles segregating in heterozygotes.

To this end, we compiled genetic information for a set of 417 mutations reported to cause fatal, recessive Mendelian diseases and estimated the frequencies of the disease-causing alleles from large exome datasets. We then compared these data to the expected frequencies of deleterious alleles based on models of MSB in order to

evaluate the effects of demography and other mechanisms in influencing these frequencies.

Results

Mendelian recessive disease allele set

We relied on two datasets, one that describes 173 autosomal recessive diseases [17] and another from a genetic testing laboratory (Counsyl; <https://www.counsyl.com/>) that includes 110 recessive diseases of clinical interest. From these lists, we obtained a set of 44 "recessive lethal" diseases associated with 42 genes (Table S1) requiring that at least one of the following conditions is met: (i) in the absence of treatment, the affected individuals die of the disease before reproductive age, (ii) reproduction is completely impaired in patients of both sexes, (iii) the phenotype includes severe mental retardation that in practice precludes reproduction, or (iv) the phenotype includes severely compromised physical development, again precluding reproduction. Based on clinical genetics datasets and the medical literature (see Methods for details), we were able to confirm that 417 Single Nucleotide Variants (SNVs) in 32 (of the 42) genes had been reported as associated with the severe form of the corresponding disease with an early-onset, and no indication of incomplete penetrance or effects in heterozygote carriers (Table S2). By this approach, we obtained a set of mutations for which, at least in principle, there is no heterozygote effect (i.e., the dominance coefficient $h = 0$ in a model with genotype fitnesses 1, 1-hs and 1-s) and the selective coefficient s for the recessive allele is 1.

A large subset of these mutations (29.3%) consists of transitions at CpG sites (henceforth CpGti), which occur at a highly elevated rates (~17-fold higher) compared to other mutation types, namely CpG transversions, and non-CpG transitions and transversions [16]. This proportion is in agreement with Cooper and Youssoufian [18], who found that ~31.5% of disease mutations in exons are CpGti.

Empirical distribution of disease alleles in Europe

Allele frequency data for the 417 variants were obtained from the Exome Aggregation Consortium (ExAC) for 60,706 individuals [19]. Out of the 417 variants associated with putative recessive lethal diseases, three were found homozygous in at least one individual in this dataset (rs35269064, p.Arg108Leu in *ASS1*; rs28933375, p.Asn252Ser in *PRF1*; and rs113857788, p.Gln1352His in *CFTR*). Available data quality information for these variants does not suggest that these genotypes are due to calling artifacts (Table S2). Since these diseases have severe symptoms that lead to early death without treatment and these ExAC individuals are purportedly healthy (i.e., do not manifest severe diseases) [19], the reported mutations are likely errors in pathogenicity classification or cases of incomplete penetrance (see a similar observation for *CFTR* and *DHCR7* in [20]). We therefore excluded them from our analyses. In addition to the mutations present in homozygotes, we also filtered out sites that had lower coverage in ExAC (see Methods), resulting in a final dataset of 385 variants in 32 genes (Table S2).

Genotypes for a subset (N = 91) of these mutations were also available for a larger sample size (76,314 individuals with self-reported European ancestry)

generated by the company Counsyl (Table S3). A comparison of the allele frequencies in this larger dataset to that of ExAC data [19] showed excellent concordance (Wilcoxon signed-rank test for paired samples, p-value=0.34; Fig S1). Thus, both data sets appear to accurately reflect European frequencies of these disease alleles, despite slight differences in the populations included. In what follows, we focus on ExAC, which includes a greater number of disease mutations.

Models of mutation-selection balance

To generate expectations for the frequencies of these disease mutations under mutation-selection balance, we considered models of infinite and finite populations of constant size [3], and conducted forward simulations using a plausible demographic model for African and European populations [21] (see Methods for details). In the models, there is a wild-type allele (A) and a deleterious allele (a, which could also represent a class of distinct deleterious alleles with the same fitness effect) at each site, such that the relative fitness of individuals of genotypes AA, Aa, or aa is given by:

- $w_{AA}=1$;
- $w_{Aa}=1-hs$;
- $w_{aa}=1-s$;

The mutation rate from A to a is u and there are no back mutations.

For a constant population of infinite size, Wright [22] showed that under these conditions, there exists a stable equilibrium between mutation and selection, when the selection pressure is sufficiently strong ($s>>u$). In particular, when the

deleterious effect of allele a is completely recessive (h=0), its equilibrium frequency q is given by:

$$q = \sqrt{u/s}. \tag{1}$$

For a *finite* population of constant size, Nei [3] derived the mean (eq. 2) and variance (eq. 3) of the frequency of a fully recessive deleterious mutation (h=0) based on a diffusion model, leading to:

$$\bar{q} = \frac{\Gamma(2Nu + 1/2)}{\sqrt{2Ns}\Gamma(2Nu)}, \tag{2}$$

$$\sigma_q^2 = u/s - \bar{q}^2, \tag{3}$$

where N is the diploid population size and Γ is the gamma function (see Simons et al. [1] for a similar approximation).

In a finite population, the mean frequency, $\bar{q}$, therefore depends on assumptions about the population mutation rate ($2Nu$). If the population mutation rate is high, such that $2Nu$>>1, $\bar{q}$ is approximated by

$$\bar{q} \approx \sqrt{u/s}, \tag{4}$$

which is independent of the population size and equal to frequency expected in an infinite size population; in particular, for a recessive lethal mutation, the mean frequency is equal to the right hand side of eq. (1). The important difference between models is that in a finite population, there is a distribution of frequencies q (because of genetic drift), rather than a single value, whose variance is given in eq. (3).

In contrast, when the finite population has a low population mutation rate ($2Nu$<<1), the mean allele frequency, $\bar{q}$, is approximated by:

$$\bar{q} \approx u\sqrt{2\pi N / s} \ ,$$

which depends on the population size. In humans, the mutation rate at each base pair is very small (on the order of 10^{-8}), so the second approximation should apply when considering each single site independently. The expectation and variance of the frequency of a fatal, fully recessive allele (i.e., s=1, h=0) are then given by:

$$\bar{q} = u\sqrt{2\pi N} \ , \tag{6}$$

and

$$\sigma_q^2 = u - \bar{q}^2 = u(1 - 2\pi Nu) \approx u \ . \tag{7}$$

All models assume complete penetrance, complete recessivity (h = 0), lethality of homozygous mutations (s = 1) and consider a single site, thereby ignoring the possibility of compound heterozygosity (unless otherwise noted).

Comparing mutation-selection balance models

Although an infinite population size has often been assumed when modeling deleterious allele frequencies (e.g. [5,23-26]), predictions under this assumption can differ markedly from what is expected from models of finite population sizes, assuming plausible parameter values for humans. For example, the long-term estimate of the effective population size from total polymorphism levels is ~20,000 individuals (assuming a per base pair mutation rate of 1.2×10^{-8} [16] and diversity levels of 0.1% [27]). In this case, the average deleterious allele frequency in the

model of finite population size is >25-fold lower than that in the infinite population model (Fig 1).

Because the human population size has not been constant, and changes in the population size can affect the frequencies of deleterious alleles in the population (as recently reviewed by Simons and Sella [2]), we also simulated mutation-selection balance under a plausible demographic model for the population history of European populations [21]. With our parameters, the mean allele frequency obtained from this model was 8.1×10^{-6}, ~2-fold higher than expected for a constant population size model with *Ne*=20,000 (Fig 1). The mean frequency seen in simulations instead matches the expectation for a constant population size of ~72k individuals (see Methods and Fig S2a). This finding is expected: the estimate of 20,000 is based on total genetic variation; assuming that most of this variation is neutral, it therefore reflects an average timescale over millions of years. For recessive lethal mutations, however, which are relatively rapidly purged by natural selection, a more recent time depth is relevant (e.g., [1]). Indeed, in our simulations, most of the disease alleles (65.6%) segregating at present arose very recently, such that they were not segregating in the population 205 generations ago, the time point after which N_e is estimated to have increased from 9,300 to 512,000 individuals [21]. Increasing the effective population size is not enough to model disease alleles appropriately however. For example, if we compare simulation results obtained under the more complex Tennessen et al. [21] demographic model [21] to those for simulations of a constant population size of N_e = 72,348, the mean allele frequencies match, but the distributions of allele frequencies are significantly different (Fig S2b).

These findings thus confirm the importance of incorporating demographic history into models for understanding the population dynamics of disease alleles [5,28,29]. In what follows, we therefore test the fit of the results based on a more realistic demographic model [21] to the observed allele frequencies.

Comparing empirical and expected distributions of disease alleles

The mutation rate, *u*, from wild-type allele to disease allele is a critical parameter in predicting the frequencies of a deleterious allele [4,30]. To model disease alleles, we considered four mutation types separately, with the goal of capturing most of the fine-scale heterogeneity in mutation rates [31-34]: transitions in methylated CpG sites (CpGti) and three less mutable types, namely transversions in CpG sites (CpGtv) and transitions and transversions outside a CpG site (nonCpGti and nonCpGtv, respectively). In order to control for the methylation status of CpG sites, which has an important influence on transition rates [31], we excluded 12 CpGti that occurred in CpG islands, which tend not to be methylated (following Moorjani et al. [35]). To allow for heterogeneity in mutation rates within each one of these four classes considered, we modeled the within-class variation in mutation rates according to a lognormal distribution (see details in Methods and [33]).

For each mutation type, we then compared results from the simulations to what is observed in ExAC, focusing on the largest sample of the same common ancestry, namely Non-Finnish Europeans (N = 33,370) (Fig 2). We find significant differences between empirical and expected mean frequencies for nonCpGti (51-fold on average; two-tailed p-value < 10^{-3}; see Methods for details) and nonCpGtv (25-

fold on average, two-tailed p-value < 10^{-3}), to a lesser extent for CpGtv (p-value = 0.05), but not for CpGti (p-value = 0.16). Intriguingly, the discrepancy between observed and expected frequencies becomes smaller as the mutation rate increases (Fig 2).

Two additional factors should further decrease the expected frequencies relative to our predictions, and will thus exacerbate the discrepancy observed between empirical and expected distribution of deleterious alleles. First, we have ignored the existence of compound heterozygosity, the case in which combinations of two distinct pathogenic alleles in the same gene lead to lethality. We know that this phenomenon is common [36], and indeed 58.4% of the 417 disease mutations considered in this study were initially identified in compound heterozygotes. Because of compound heterozygosity, each deleterious mutation will be selected against not only when present in two copies within the same individual, but also in the presence of certain lethal mutations at other sites of the same gene. Thus, the frequency of a deleterious mutation will be lower under a model incorporating compound heterozygosity than under a model that does not include this phenomenon.

In order to model the effect of compound heterozygosity in our simulations, we reran our simulations considering a gene rather than a site. In these simulations, we used the same setup as in the site level analysis, except for the mutation rate, here defined as U, the sum of the mutation rates u_i at each site i that is known to cause a severe and early onset form of the disease (Table S2; see Methods for details). This approach does not consider the contribution of other mutations in the

genes that cause the mild and/or late onset forms of the disease, and implicitly assumes that all combinations of known recessive lethal alleles of the same gene have the same fitness effect as homozygotes. Comparing observed frequencies to those generated by simulation, one third of the genes differ at the 5% level, with a clear trend for empirical frequencies to be above expectation (Table S4; Fig 3; Fisher's combined probability test p-value < 10^{-14}).

This finding is even more surprising than it may seem, because we do not know the complete mutation target for each gene and are therefore ignoring the contribution of additional sites at which disease mutations could arise. If there are undiscovered recessive lethal mutations that cause death when forming compound heterozygotes with the ascertained ones, the purging effect of purifying selection on the known mutations will be under-estimated, leading us to over-estimate the expected frequencies in simulations. Therefore, our predictions are, if anything, an over-estimate of the expected allele frequency and the discrepancy between predicted and the observed is likely even larger than what we found.

The other factor that we did not consider in simulations but would reduce the expected allele frequencies is a subtle fitness decrease in heterozygotes. To evaluate potential fitness effects in heterozygotes when none had been documented in humans, we considered the phenotypic consequences of orthologous gene knockouts in mouse. We were able to retrieve information on phenotypes for both homozygote and heterozygote mice for only eight out of the 32 genes, namely *ASS1, CFTR, DHCR7, NPC1, POLG, PRF1, SLC22A5,* and *SMPD1*. For all eight, homozygote knockout mice presented similar phenotypes as affected humans, and heterozygotes

showed a milder but detectable phenotype (Table S5). The extent to which the same phenomenon applies to the mutations with no clinically reported effects in heterozygous humans is unclear, but the finding with knockout mice makes it plausible that there exists a very small fitness decrease in heterozygotes in humans as well, potentially enough to have a marked impact on the allele frequencies of deleterious mutations, but not enough to have been recognized in clinical investigations. Indeed, even if the fitness effect in heterozygotes were h = 1%, a 68% decrease in the mean allele frequency of the disease allele is expected (Fig S3).

Discussion

To investigate the population genetics of human disease, we focused on mutations that cause Mendelian, recessive disorders that lead to early death or completely impaired reproduction. We sought to understand to what extent the frequencies of these mutations fit the expectation based on a simple balance between the input of mutations and the purging by purifying selection, as well as how other mechanisms might affect these frequencies. Many studies implicitly or explicitly compare known disease allele frequencies to expectations from mutation-selection balance. We tested whether known disease alleles as a class fit these expectations, and found that, under a sensible demographic model for European population history with purifying selection only in homozygotes, the expectations fit the observed disease allele frequencies well when the mutation rate is high (see CpGti in Fig 2). If mutation rate is low, however, as is the case for CpGtv, nonCpGti and nonCpGtv, the mean empirical frequencies of disease alleles are above

expectation (Fig 2). Further, including possible effects of the disease in heterozygote carriers or for the effect of compound heterozygotes in these models only increases the discrepancy.

In principle, higher than expected disease allele frequencies could be explained by at least four (non-mutually exclusive) factors: (i) misspecification of the demographic model, (ii) widespread errors in reporting the causal variant, (iii) overdominance of disease alleles and (iv) low penetrance of disease mutations. Notably, it has been estimated that population growth in Europe could have been stronger that we considered in our simulations [37,38]. Stronger population growth does increase the expected frequency of recessive disease alleles (Fig S4, columns A-D). However, the impact of more intense growth is likely insufficient to explain the observed discrepancy: the allele frequencies observed in ExAC are still on average an order of magnitude larger than expected based on a model with ten-fold more intense growth than in [21] (Fig S4). Similarly, while errors in reporting the causal variants are known to be common [19,39,40], we attempted to minimize them by filtering out any case for which there was not compelling evidence of association with a recessive lethal disease, reducing our initial set of 769 mutations to 385 in which we had greater confidence (see Methods for details).

The other two factors, overdominance and low penetrance, are likely explanations for a subset of cases. For instance, *CFTR*, the gene in which mutations lead to cystic fibrosis, is the furthest above expectation (p-value = 0.002; Fig 3). It was long noted that there is an unusually high frequency of the *CFTR* deletion ΔF508 in Europeans, which led to speculation that disease alleles of this gene may be

subject to over-dominance ([41], but see [42]). Whether or not this is the case, there is evidence that disease mutations in this gene can complement one another [8,9] and that modifier loci in other genes also influence their penetrance [9,12]. Consistent with variable penetrance, Chen et al. [20] identified three purportedly healthy individuals carrying two copies of disease mutations in this gene. Similarly, *DHCR7*, the gene associated with the Smith-Lemli-Opitz syndrome, is somewhat above expectation in our analysis (p-value = 0.052; Fig 3) and healthy individuals were found to be homozygous carriers of putatively lethal disease alleles in other studies [20]. These observations make it plausible that, in a subset of cases (particularly for *CFTR*), the high frequency of deleterious mutations associated with recessive, lethal diseases are due to genetic interactions that modify the penetrance of certain recessive disease mutations.

Nonetheless, two factors argue against modifiers alleles being a sufficient explanation for the site-level analysis. First, when we remove 130 mutations in *CFTR* and 12 in *DHCR7*, two genes that were outliers at the gene-level (Fig 3; Table S4) and for which healthy individuals were reported to be homozygous for a deleterious allele [20], the discrepancy between observed and expected allele frequencies is barely impacted (Fig S5). Second, there is no obvious reason why this explanation would apply differently to distinct types of mutations, yet the extent to which observed allele frequencies exceed the expected depends on the mutation rate, with no departure seen at CpGti (Fig 2).

Instead, it seems plausible that there is likely an ascertainment bias in disease allele discovery and mutation identification. Since not all mutations that can cause a

specific Mendelian disease are known, those mutations that were discovered to date are likely the ones that by chance have drifted to higher frequencies, and are thus more likely to have been mapped and be reported in the medical literature. One clear implication of that is that there are numerous sites at which mutations cause recessive lethal diseases yet to be discovered, particular at non-CpG sites. More generally, this ascertainment bias complicates the interpretation of observed allele frequencies in terms of the selection pressures acting on disease alleles.

Beyond this specific point, our study illustrates how the large sample sizes now made available to researchers in the context of projects like ExAC [19] can be used not only for direct discovery of disease variants, but also to test why there are disease alleles segregating in the population and to understand at what frequencies we might expect to find them.

Methods

Disease allele set

In order to identify single nucleotide variants within the 42 genes associated with lethal, recessive Mendelian diseases (Table S1), we initially relied on the ClinVar dataset [43] (accessed on June 3rd, 2015). We filtered out any variant that is an indel or a more complex copy number variant or that is ever classified as benign or likely benign in ClinVar (whether or not it is also classified as pathogenic or likely pathogenic). By this approach, we obtained 769 SNVs described as pathogenic or likely pathogenic. For each one of these variants, we searched the literature for evidence that it is exclusively associated to the lethal and early onset form of the

disease and was never reported as causing the mild and/or late-onset form of the disease. We considered effects in the absence of medical treatment, as we were interested in the selection pressures acting on the alleles over evolutionary scales rather than in the last one or two generations. Variants with mention of incomplete penetrance (i.e. for which homozygotes were not always affected) or with known effects in heterozygote carriers were removed from the analysis. This process yielded 417 SNVs in 32 genes associated with distinct Mendelian recessive lethal disorders (Table S2). Although these mutations were purportedly associated with complete recessive diseases, we sought to examine whether there would be possible, unreported effects in heterozygous carriers. To this end, we used the Mouse Genome Database (MGD) [44] (accessed July 29th, 2015) and were able to retrieve information for both homozygote and heterozygote mice for eight out of the 32 genes (all of which with a homologue in mice) (Table S5).

In addition to the information provided by ClinVar for each one of these variants, we considered the immediate sequence context of each SNV, to tailor the mutation rate estimate accordingly [16]. For doing so, we used an in-house Python script and the human genome reference sequence hg19 from UCSC (<http://hgdownload.soe.ucsc.edu/goldenPath/hg19/chromosomes/>).

Genetic datasets

The Exome Aggregation Consortium (ExAC) [19] was accessed on August 9th, 2016. The data consist of genotype frequencies for 60,706 individuals, assigned

after PCA analysis to one of seven population labels: African (N=10,406), East Asian (N=8,654), Finnish (N=6,614), Latino (N=11,578), Non-Finnish European (N=66,740), South Asian (N=16,512) and "other" (N=908). We focused our analyses on those individuals of Non-Finnish European descent, because they constitute the largest sample size from a single ancestry group. We note that, some diseases mutations, for instance, those in *ASPA*, *HEXA* and *SMPD1*, are known to be especially prevalent in Ashkenazi Jewish populations, and that could potentially bias our results if Ashkenazi Jewish individuals constituted a great portion of the sample we considered. However, this sample includes only ~2,000 (~3%) Ashkenazi individuals (Dr. Daniel MacArthur, personal communication).

From the initial 417 mutations, we filtered out three that were homozygous in at least one individual in ExAC and 29 that had lower coverage, i.e., fewer than 80% of the individuals were sequenced to at least 15x. This approach left us with a set of 385 mutations with a minimum coverage of 27x per sample and an average coverage of 69x per sample (Table S2). For 248 sites with non-zero sample frequencies, ExAC reported the number of non-Finnish European individuals that were sequenced, which was on average 32,881 individuals [19]. For the remaining 137 sites, we did not have this information. Nonetheless, the mean coverage is reported for all sites and does not differ between the two sets of sites (Fig S6). We therefore assumed that mean number of individuals covered for all sites was 32,881 [45] and used this number to obtain sample frequencies from simulations, as explained below.

A second genetic dataset was obtained from Counsyl (<https://www.counsyl.com/>). Counsyl is a commercial genetic screening laboratory that offers, among other products, the "Family Prep Screen", a genetic screening test intended to detect carrier status for up to 110 recessive Mendelian diseases in couples that are planning to have a child. A subset of 294,000 of its customers was surveyed by genotyping or sequencing for "routine carrier screening". This subset excludes individuals with indications for testing because of known personal or family history of Mendelian diseases, infertility, and consanguinity. It therefore represents a more random (with regard to the presence of disease alleles), population-based survey. For these individuals, we had details on self-reported ancestry (14 distinct ethnic/ancestry/geographic groups) and the allele frequencies for 98 mutations that match those that passed our variant selection criteria described above, of which 92 are also sequenced to high coverage in the ExAC database (Table S2). We focused our analysis of this dataset on the 76,314 individuals with self-reported Northern or Southern European ancestry.

Simulating the evolution of disease alleles with population size change

We also modeled the frequency of a deleterious allele in human populations by forward simulations based on a crude but plausible demographic model for human populations from Africa and Europe, inferred from exome data for African-Americans and European-Americans [21]. To this end, we used a program described in [1]. In brief, the demographic scenario consists of an Out-of-Africa demographic model, with changes in population size throughout the population history, including

a severe bottleneck in Europeans following the split from the African population and a rapid, recent population growth in both populations [21]. As in Simons et al. [1], we simulated genetic drift and two-way gene flow between Africans and Europeans in recent history. Negative selection acting on a single bi-allelic site was modeled as in the analytic models.

Allele frequencies follow a Wright-Fisher sampling scheme in each generation according to these viabilities, with migration rate and population sizes varying according to the demographic scenario considered. Whenever a demographic event (e.g. growth) altered the number of individuals and the resulting number was not an integer, we rounded it to the nearest integer, as in Simons et al. [1]. A burn-in period of 10*Ne* generations with constant population size *Ne* = 7,310 individuals was implemented in order to ensure an equilibrium distribution of segregating alleles at the onset of demographic changes in Africa, 148 Kya.

In contrast to Simons et al. [1], our simulations always start with the ancestral allele A fixed and mutation occurs exclusively from this allele to the deleterious one (a), i.e. a mutation occurs with mean probability *u* per gamete, per generation, and there is no back-mutation. However, recurrent mutations at a site are allowed, as in Simons et al. [1].

When implementing the model, we used mean mutation rates *u* estimated from a large human pedigree study [16], considering the genomic average (1.2×10^{-8} per base pair, per generation) and four distinct mutation types (CpGti = 1.12×10^{-7}; CpGtv = 9.59×10^{-9}; nonCpGti = 6.18×10^{-9}; and nonCpGtv = 3.76×10^{-9}). While these four categories capture much of the variation in germline mutation rates across

sites, a number of other factors (e.g., the larger sequence context or the replication timing) also influence mutation rates, introducing heterogeneity in the mutation rate within each class considered [31-33,46]. To allow for this heterogeneity as well as for uncertainty in the point mutation rates estimates, in each simulation, instead of using a fixed rate u for each mutation type, we drew the mutation rate M from a lognormal distribution with the following parameters:

$$\log_{10} M \mid u \sim N(\log_{10} u - \frac{\sigma^2}{2}\ln(10), \sigma^2) \tag{7}$$

such that that $E[M]=u$. σ was set to 0.57 (following [33]).

By this procedure, we ran two million simulations for each mutation type, thus obtaining the distribution of deleterious allele frequencies expected for the European population. In order to compare simulation results to the empirical data, we subsampled the simulated population to match the average number of autosomal chromosomes in the non-Finnish European sample from ExAC (N = 65,762 chromosomes).

To measure the significance of the deviation between observed and expected allele frequencies, we proceeded as follows: First, we sampled K allele frequencies from the 2M simulations implemented for each mutation type, where K *is* the number of mutations described for that type. We repeated this step 1,000 times, thus obtaining a distribution for the mean allele frequency across K mutations. Finally, to obtain a two-tailed p-value, we considered the rank of the empirical mean relative to simulated outcomes.

A well-known source of heterogeneity in mutation rate within the CpGti class is methylation status, with a high transition rate seen only at methylated CpGs [18]. In our analyses, we tried to control for the methylation status of CpG sites by excluding sites located in CpG islands (CGIs). The CGI annotation for hg19 was obtained from UCSC Genome Browser (track "Unmasked CpG"; <http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/cpgIslandExtUnmasked.txt.gz>, accessed in June 6th, 2016) and BEDTools [47] was used to exclude those CpG sites located in CGIs. We note that the CpGti estimate from [16] includes CGIs, and in that sense the average mutation rate that we are using for CpGti may be a very slight underestimate of the mean rate for transitions at methylated CpG sites.

Unless otherwise noted, the expectation assumes fully recessive, lethal alleles with complete penetrance. Notably, by calculating the expected frequency one site at a time, we are ignoring possible interaction between genes (i.e., effects of the genetic background) and among different mutations within a gene (i.e., compound heterozygotes). These assumptions are relaxed in two ways. In one analysis (Fig S3), we consider a very low selective effect in heterozygous individuals (h = 1%), reasoning that such an effect could plausibly go undetected in medical examinations and yet would nonetheless impact the frequency of the disease allele. Second, when considering the gene-level analysis (Fig 3), we implicitly allow for compound heterozygosity between any pair of lethal mutations. For this analysis, we ran 1000 simulations for a total mutation rate U per gene that was calculated accounting for the heterogeneity and uncertainty in the mutation rates estimates as follows: (i) For sites known to cause a recessive lethal disease and that passed our filtering criteria

(Table S2), we drew a mutation rate u_i from the lognormal distribution, as described above; (ii) We then took the sum of u_i as the total mutation rate U; (iii) We then ran one replicate with U as the mutation parameter, and other parameters as specified for site level analysis. Because the mutational target size considered in simulations is only comprised of those sites at which mutations are known to cause a lethal recessive disease, it is almost certainly an underestimate of the true mutation rate—potentially by a lot. We note further that by this approach, we are assuming that compound heterozygotes formed by any two lethal alleles have fitness zero, i.e., that they are identical in their effects to homozygotes for any of the lethal alleles. Moreover, we are implicitly ignoring the possibility of complementation, which is (somewhat) justified by our focus on mutations with severe effects and complete penetrance (but see Discussion). Since we were interested in understanding the effect of compound heterozygosity, for this analysis, we did not consider the five genes in which only one mutation passed our filters (*BCS1L, FKTN, LAMA3, PLA3G6*, and *TCIRG1*).

All codes and data to generate the figures in R [48] and the script used to get the sequence context of each mutation (kindly provided by Ellen Leffler) are available at https://github.com/cegamorim/PopGenHumDisease. The code to run the simulations is available at https://github.com/sellalab/PopGenHumDisease. Allele frequencies and other information for the disease mutations employed in the analyses are in Tables S2 and S3.

Acknowledgements

We thank Dr. Daniel MacArthur for providing information on accessing ExAC information. CEGA was partially funded by a Science Without Borders fellowship from CAPES foundation (BEX 8279/11-0) and CNPq (PDE 201145/2015-4), Brazil. ZG was partially supported by a postdoctoral fellowship funded by Stanford Center for Computational, Evolutionary and Human Genomics. JFD was funded by a Science Without Borders fellowship from CAPES foundation (88888.038761/2013-00). The work was partially supported by a Research Initiative in Science and Engineering grant from Columbia University to JKP and MP. The computing in this project was supported by two National Institutes of Health instrumentation grants (S10OD012351 and S10OD021764) received by the Department of Systems Biology at Columbia University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceived the study: JP, MP. Designed the study: CEGA, ZG, MP. Analyzed the data: CEGA. Implemented analytical models: ZG. Wrote the paper: CEGA, ZG, MP. Helped in acquisition and analysis of data: ZB, JFD. Contributed analytical tools or data: YBS, IR.

References

1. Simons YB, Turchin MC, Pritchard JK, Sella G (2014) The deleterious mutation load is insensitive to recent population history. Nat Genet 46: 220-224.
2. Simons YB, Sella G (2016) The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. bioRxiv.
3. Nei M (1968) The frequency distribution of lethal chromosomes in finite populations. Proc Natl Acad Sci U S A 60: 517-524.
4. Gillespie JH (2004) Population Genetics: A Concise Guide. Baltimore, MD: Johns Hopkins University Press.
5. Brandvain Y, Wright SI (2016) The Limits of Natural Selection in a Nonequilibrium World. Trends Genet 32: 201-210.
6. Balick DJ, Do R, Cassa CA, Reich D, Sunyaev SR (2015) Dominance of Deleterious Alleles Controls the Response to a Population Bottleneck.
7. Beauchamp KA, Muzzey D, Wong KK, Hogan GJ, Karimi K, et al. (2016) Systematic Design and Comparison of Expanded Carrier Screening Panels. bioRxiv.
8. Cormet-Boyaka E, Jablonsky M, Naren AP, Jackson PL, Muccio DD, et al. (2004) Rescuing cystic fibrosis transmembrane conductance regulator (CFTR)-processing mutants by transcomplementation. Proc Natl Acad Sci U S A 101: 8221-8226.
9. Rapino D, Sabirzhanova I, Lopes-Pacheco M, Grover R, Guggino WB, et al. (2015) Rescue of NBD2 mutants N1303K and S1235R of CFTR by small-molecule correctors and transcomplementation. PLoS One 10: e0119796.
10. Andressoo JO, Jans J, de Wit J, Coin F, Hoogstraten D, et al. (2006) Rescue of progeria in trichothiodystrophy by homozygous lethal Xpd alleles. PLoS Biol 4: e322.
11. Gallati S (2014) Disease-modifying genes and monogenic disorders: experience in cystic fibrosis. Appl Clin Genet 7: 133-146.
12. Corvol H, Blackman SM, Boelle PY, Gallins PJ, Pace RG, et al. (2015) Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. Nat Commun 6: 8382.
13. Habara A, Steinberg MH (2016) Minireview: Genetic basis of heterogeneity and severity in sickle cell disease. Exp Biol Med (Maywood) 241: 689-696.
14. Hedrick PW (2011) Population genetics of malaria resistance in humans. Heredity (Edinb) 107: 283-304.
15. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, et al. (2015) The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet 97: 199-215.
16. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. Nature 488: 471-475.
17. Turner TN, Douville C, Kim D, Stenson PD, Cooper DN, et al. (2015) Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense mutation distribution patterns. Hum Mol Genet.

18. Cooper DN, Youssoufian H (1988) The CpG dinucleotide and human genetic disease. Hum Genet 78: 151-155.
19. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536: 285-291.
20. Chen R, Shi L, Hakenberg J, Naughton B, Sklar P, et al. (2016) Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. Nat Biotechnol 34: 531-538.
21. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337: 64-69.
22. Wright S (1937) The Distribution of Gene Frequencies in Populations. Proc Natl Acad Sci U S A 23: 307-320.
23. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17: 502-510.
24. Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet 11: 2417-2423.
25. Zwick ME, Cutler DJ, Chakravarti A (2000) Patterns of genetic variation in Mendelian and complex traits. Annu Rev Genomics Hum Genet 1: 387-407.
26. Cassa CA, Weghorn D, Balick DJ, Jordan DM, Nusinow D, et al. (2016) Estimating the Selective Effect of Heterozygous Protein Truncating Variants from Human Exome Data. bioRxiv.
27. The Genomes Project C (2015) A global reference for human genetic variation. Nature 526: 68-74.
28. Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, et al. (2008) Proportionally more deleterious genetic variation in European than in African populations. Nature 451: 994-997.
29. Peischl S, Dupanloup I, Bosshard L, Excoffier L (2016) Genetic surfing in human populations: from genes to genomes. Curr Opin Genet Dev 41: 53-61.
30. Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant...or not?
31. Segurel L, Wyman MJ, Przeworski M (2014) Determinants of mutation rate variation in the human germline. Annu Rev Genomics Hum Genet 15: 47-70.
32. Aggarwala V, Voight BF (2016) An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet 48: 349-355.
33. Harpak A, Bhaskar A, Pritchard JK (2016) Effects of variable mutation rates and epistasis on the distribution of allele frequencies in humans. bioRxiv.
34. Hodgkinson A, Eyre-Walker A Variation in the mutation rate across mammalian genomes.
35. Moorjani P, Amorim CE, Arndt PF, Przeworski M (2016) Variation in the molecular clock of primates. Proc Natl Acad Sci U S A 113: 10607-10612.
36. Kamphans T, Sabri P, Zhu N, Heinrich V, Mundlos S, et al. (2013) Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. PLoS One 8: e70151.

37. Gazave E, Ma L, Chang D, Coventry A, Gao F, et al. (2014) Neutral genomic regions refine models of recent rapid human population growth. Proc Natl Acad Sci U S A 111: 757-762.
38. Gao F, Keinan A (2016) Explosive genetic evidence for explosive human population growth. Curr Opin Genet Dev 41: 130-139.
39. Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, et al. (2012) Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am J Hum Genet 91: 1022-1032.
40. Piton A, Redin C, Mandel JL (2013) XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. Am J Hum Genet 93: 368-383.
41. Gabriel SE, Brigman KN, Koller BH, Boucher RC, Stutts MJ (1994) Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. Science 266: 107-109.
42. Quinton PM (1994) Human genetics. What is good about cystic fibrosis? Curr Biol 4: 742-743.
43. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, et al. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research 42: D980-D985.
44. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE (2015) The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic Acids Res 43: D726-736.
45. Haque IS, Lazarin GA, Kang H, Evans EA, Goldberg JD, et al. (2016) MOdeled fetal risk of genetic diseases identified by expanded carrier screening. JAMA 316: 734-742.
46. Mugal CF, Ellegren H (2011) Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. Genome Biol 12: R58.
47. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.
48. R Core Team (2015) R: A Language and Environment for Statistical Computing. Vienna, Austria.
49. Online Mendelian Inheritance in Man O (2016). Baltimore, MD: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University.

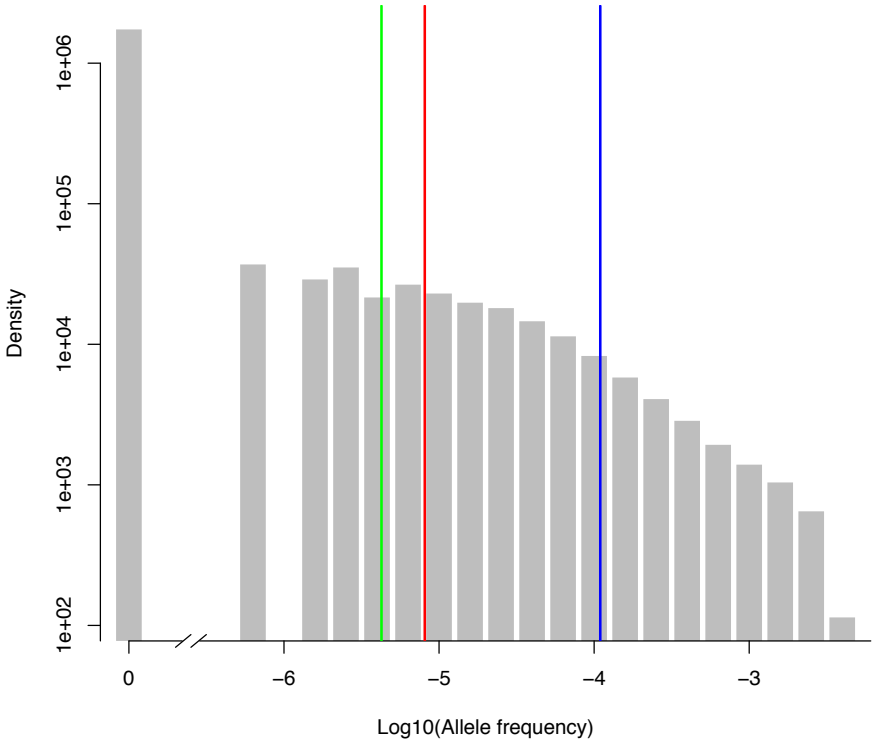


Fig 1. Expected allele frequencies in the population, under a mutation-selection balance models. Blue bar denotes the expected under an infinite population size, a green bar under a finite constant population, and a red bar under a plausible demographic model for European populations (distribution shown in the grey histogram). All models assume s=1 and h=0. For the finite constant population size model, we present the mean frequency expected for a population size of 20,000 (see Fig S2a for the effect of varying the population size). Sample allele frequencies (q) were transformed to log10(q) and those q=0 were set to 10^{-7} for visual purposes, but indicated as "0" on the X-axis. Y-axis is plotted on a log-scale.

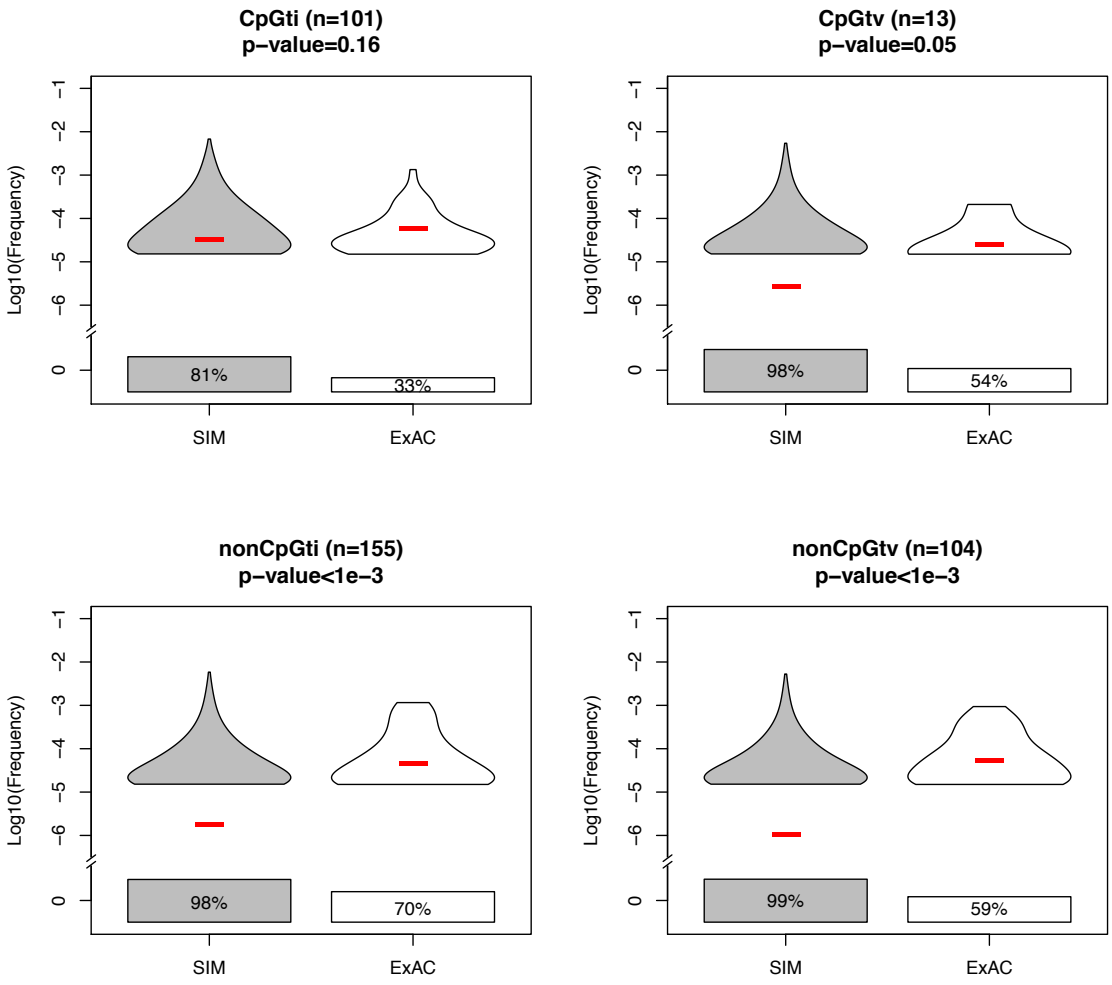


Fig 2. Expected and observed distributions of disease allele frequencies. The four panels correspond to four different mutation types. The title of the panel indicates the mutation type, followed by the total number of mutations, with p-value for the difference between observed and expected mean frequencies below it. Results in grey (SIM) were obtained from simulations considering a plausible demographic model for European populations [21] (see text). The observed values estimated from 33,370 individuals of European ancestry from ExAC are shown in white. Violin plots show the density distribution of variable sites, while boxes indicate the proportion of sites for which the wild, non-deleterious mutation is fixed. Means (considering both segregating and fixed mutations) are indicated with red horizontal bars.

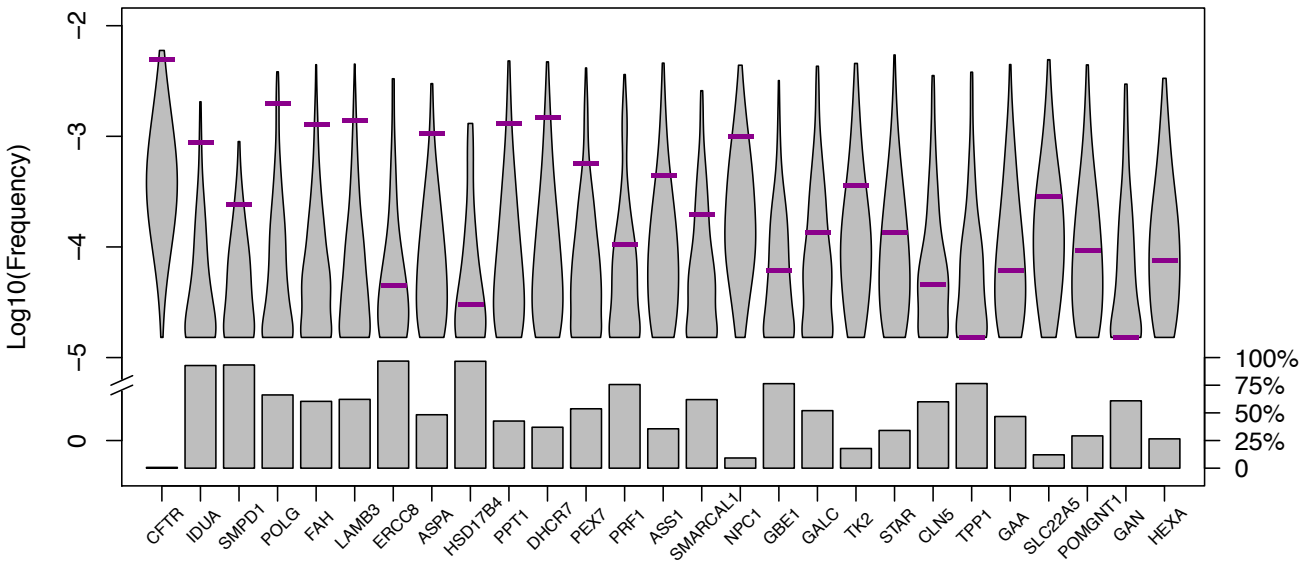


Fig S3. Disease allele frequencies at the gene level. The expectation (grey) is based on 1000 simulations, assuming no effect in heterozygotes, but allowing for compound heterozygosity (see Methods for details). Observed frequencies (purple bars) were obtained from ExAC considering 33,370 European individuals. Genes are ordered according to the two-tailed p-value for the significance of the deviation of the empirical frequencies from the expected (Table S4). To calculate that, we considered the rank of the empirical mean relative to the simulations. Violin plots show the distribution of simulated allele frequencies among segregating alleles and boxes represent the fraction of simulations in which no deleterious allele was observed in the simulated sample at present.

Table S5. Phenotypic effect of mouse knock-outs (see main text)

Gene	Human disease	OMIM number	Phenotype of affected human cases[a]	Phenotype of homozygous knockout mice[b]	Phenotype of heterozygous knockout mice[b]
ASS1	Citrullinemia	215700	Very high concentration of the amino-acid citrulline in serum, spinal fluid, and urine.	Complete neonatal lethality, abnormal circulating amino-acid level, increased circulating ammonia level.	Abnormal circulating amino-acid level.
CFTR	Cystic fibrosis	219700	Disruption of exocrine function of the pancreas, intestinal glands (meconium ileus), biliary tree (biliary cirrhosis), bronchial glands (chronic bronchopulmonary infection with emphysema) and sweat glands (high sweat electrolyte with depletion in a hot environment). Infertility occurs in males and females.	Partial postnatal lethality, aphagia, pancreatic acinar cell atrophy, abnormal intestine morphology, abnormal digestive system physiology, abnormal gland morphology, acute pancreas inflammation, weight loss, distended abdomen, abnormal ion homeostasis, enlarged gallbladder, abnormal respiratory system physiology, lacrimal gland atrophy.	Impaired fertilization, decreased litter size.
DHCR7	Smith-Lemli-Opitz syndrome	270400	Multiple congenital malformation and mental retardation syndrome.	Complete neonatal lethality, abnormal suckling behavior, weakness, abnormal nasal cavity morphology, fetal growth retardation, cyanosis, abnormal brain development, distended urinary bladder.	Abnormal cholesterol level, syndactyly, partial embryonic lethality, decreased brain size.
NPC1	Niemann-Pick disease, type C1	257220	Lipid storage disorder characterized by progressive neurodegeneration.	Premature death, abnormal Purkinje cell morphology, increased brain cholesterol level, increased liver cholesterol level, abnormal macrophage morphology, abnormal microglial cell activation, abnormal lipid homeostasis, decreased body weight,	Increased brain cholesterol level.

				impaired coordination.	
POLG	Alpers syndrome	203700	Clinical triad of psychomotor retardation, intractable epilepsy, and liver failure in infants and young children. Pathologic findings include neuronal loss in the cerebral gray matter with reactive astrocytosis and liver cirrhosis.	Premature death, abnormal mitochondrial physiology, decreased thymocyte number, abnormal lymphopoiesis, macrocytic anemia, abnormal erythroid lineage cell morphology.	Abnormal bone marrow cell physiology, increased B cell derived lymphoma incidence.
PRF1	Hemophagocytic lymphohistiocytosis	603553	Immune dysregulation characterized clinically by fever, edema, hepatosplenomegaly, and liver dysfunction. Neurologic impairment, seizures, and ataxia are frequent.	Increased activated T cell number, decreased cytotoxic T cell cytolysis, abnormal cytokine secretion, decreased susceptibility to autoimmune diabetes, increased susceptibility to viral infection, premature death, complete postnatal lethality, liver inflammation, CNS inflammation, abnormal circulating cytokine level, decreased leukocyte cell number.	Insulitis, periinsulitis, impaired natural killer cell mediated cytotoxicity.
SLC22A5	Carnitine deficiency	212140	This results in impaired fatty acid oxidation in skeletal and heart muscle. In addition, renal wasting of carnitine results in low serum levels and diminished hepatic uptake of carnitine by passive diffusion, which impairs ketogenesis.	Premature death, enlarged liver, hepatic steatosis, increased triglyceride level, decreased circulating carnitine level, impaired lipolysis, decreased body weight, enlarged heart.	Decreased circulating carnitine level, impaired lipolysis.
SMPD1	Niemann-Pick disease, type A	257200	The clinical phenotype for type A ranges from a severe infantile form with neurologic degeneration resulting in death usually by 3 years of age.	Premature death, ataxia, lethargy, abnormal apoptosis, decreased body weight, increased macrophage derived foam cell number, abnormal lipid homeostasis, increased susceptibility to bacterial	Abnormal immune system cell morphology, abnormal neuron differentiation, abnormal depression-related behavior.

	infection, decreased brain size.

Phenotypes obtained from [49][a] and [44][b]

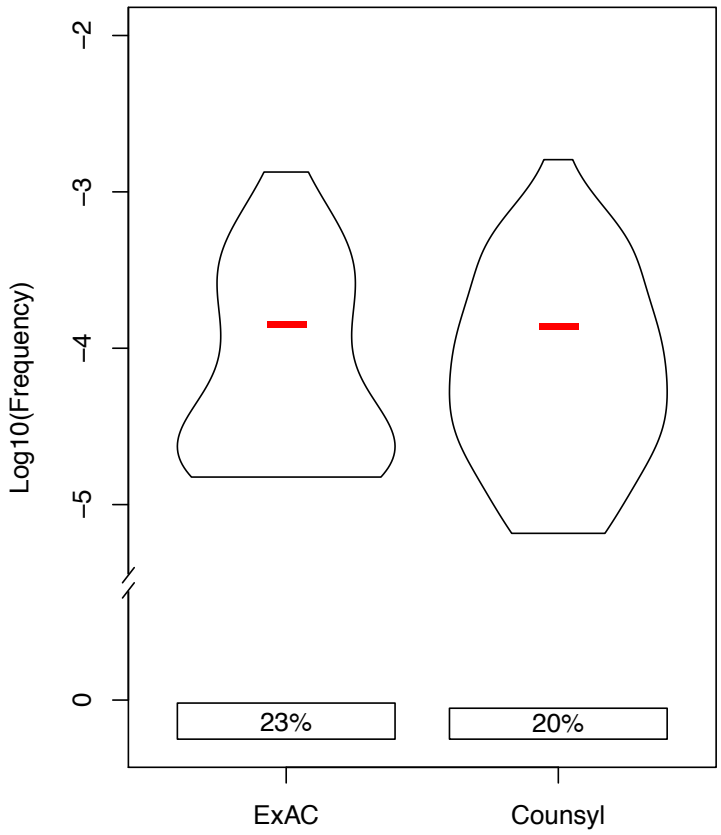
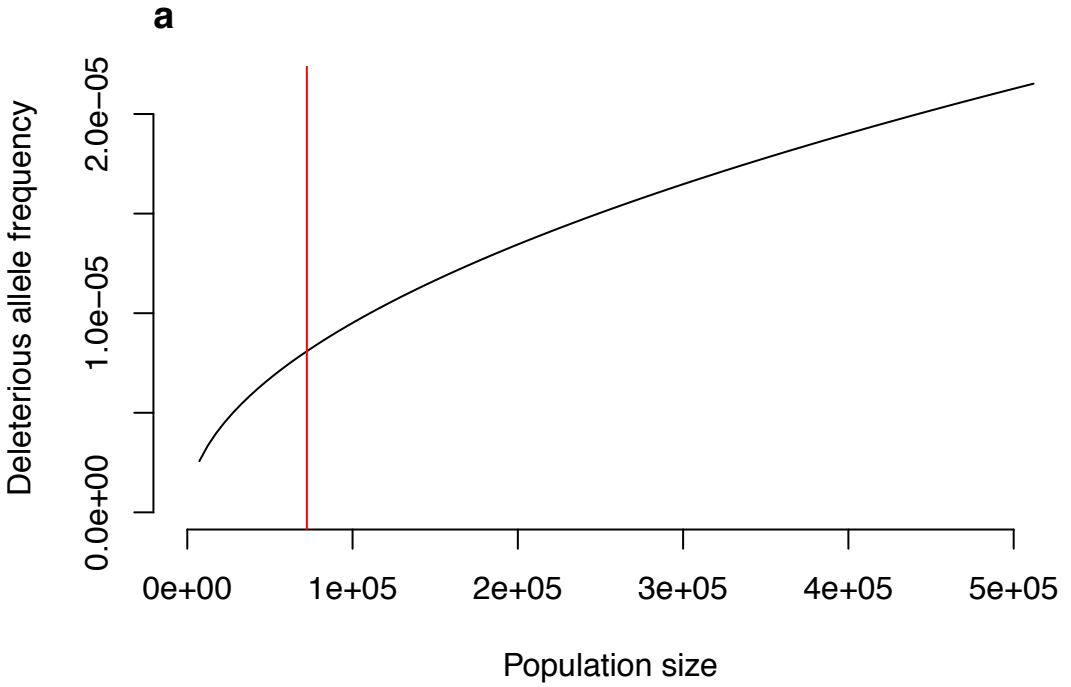


Fig S1. Frequency distribution of disease mutations in individuals of European ancestry. Shown are allele frequencies for 92 variants associated with lethal, recessive diseases, as estimated from 33,370 individuals non-Finnish, European-ancestry in the Exome Aggregation Consortium (ExAC) database [19] and 76,314 European-ancestry individuals from a genetic testing laboratory (Counsyl) (see Methods). Allele frequencies do not differ significantly between datasets (Wilcoxon signed-rank test for paired samples, p-value=0.34). Violin plots show the density distribution of alleles segregating in these samples, while boxes indicate the proportion of sites for which the deleterious mutation was not observed.


a
Deleterious allele frequency
0.0e+00
1.0e−05
2.0e−05
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Population size

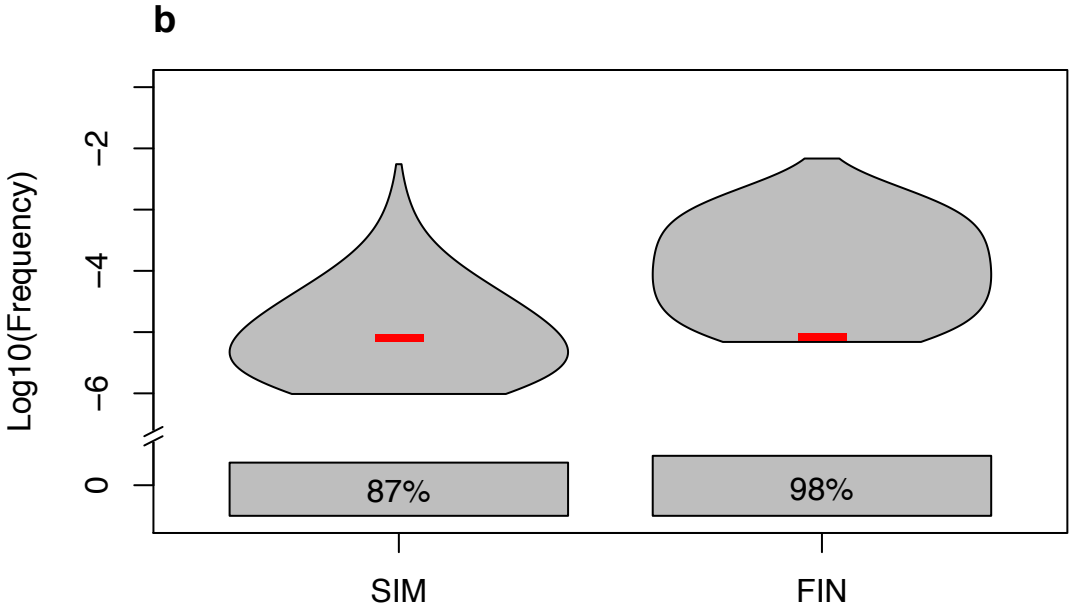

b
Log10(Frequency)
−2
−4
−6
0
87%
98%
SIM
FIN

Fig S2. Comparisons of SIM and FIN models. (A) Mean allele frequency as a function of effective population size, under the FIN model. The X-axis range corresponds to the minimum and maximum effective population size estimated in [21]. The red bar indicates the value of a constant population size at which the mean allele frequency predicted under a constant population size model is the same as the mean allele frequency estimated in simulations (SIM), for an average mutation rate of 1.20×10^{-8} [16]. Even for this value, the mean matches the constant size expectation, but the distributions differ (see panel "b" below). (B) SIM refers to the complex demographic scenario inferred by Tennessen et al. [21] for the evolution of European populations (see Methods). In the finite, constant size population model (FIN), N is set to 72,348 individuals, so that the mean allele frequency (red bars) is the same as in simulations (see panel "a" above). To generate a distribution for this model, we simulated a constant population size, as described in Methods for the Tennessen et al. model. Violin plots show the distribution of allele frequencies among segregating alleles, while the boxes indicate the proportion of sites at which the non-disease allele is fixed. These models assume strong selection (s=1) and complete recessivity (h=0). As can be seen, the two distributions differ significantly (Kolmogorov-Smirnov test, p-value $< 10^{-15}$).

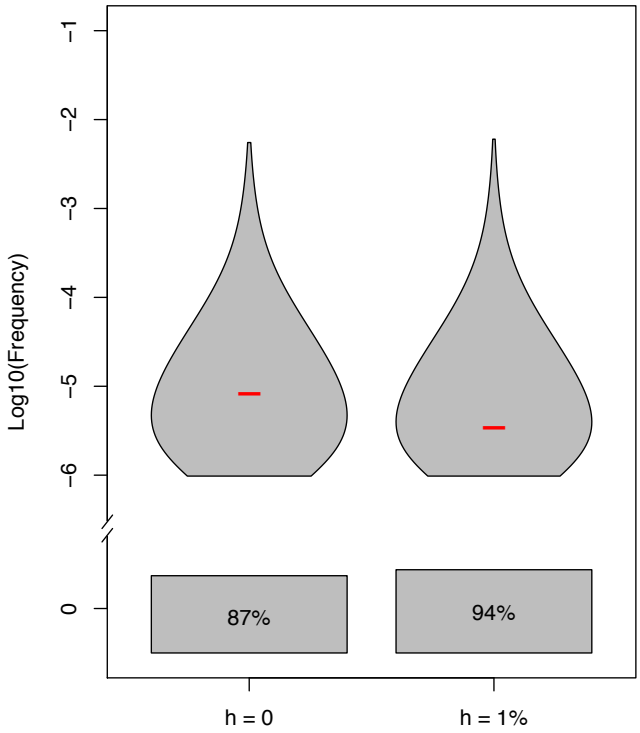


Fig S3. The impact on disease allele frequencies of a small fitness effect in heterozygotes (*h*=0.01). Shown is the distribution generated from simulations. Means are represented by red horizontal bars. Violin plots show the density distribution of alleles segregating in these samples, whereas boxes indicate the proportion of sites for which the deleterious mutation was not observed. When a small fitness effect in heterozygotes is considered in the simulations, the mean decreases by 68% and a larger proportion of sites are not segregating. The two distributions differ significantly (p-value < 10^{-15} by a Kolmogorov-Smirnov test).

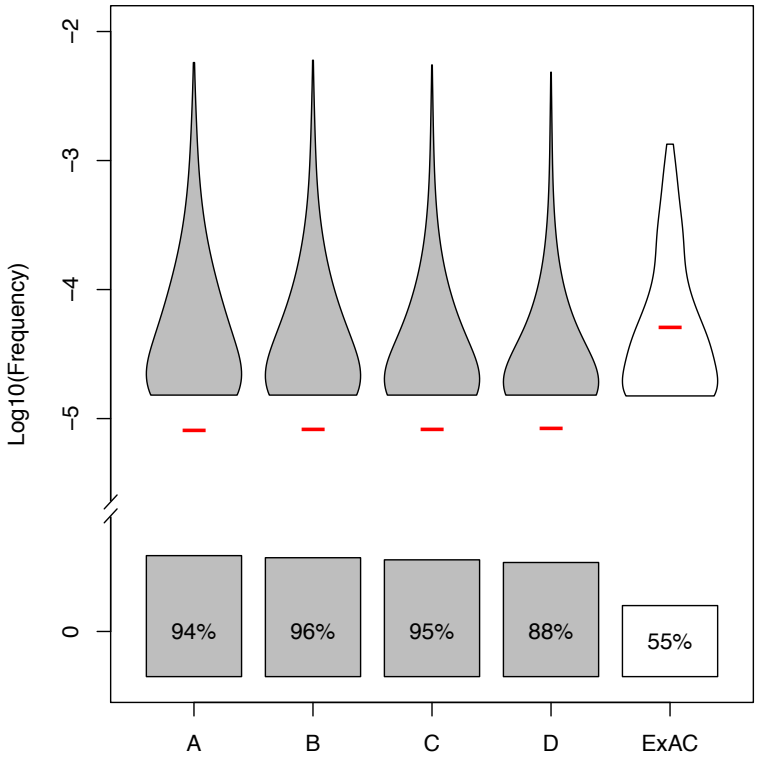


Fig S4. Effect of varying the end population size in the SIM model. Tennessen et al. [21] inferred the present effective population size of Europeans to be 512,000 individuals (column A). We considered the effect of larger population sizes (2-, 4-, and 10-fold increase, denoted by columns B, C and D respectively), keeping other parameters the same. The observed allele frequency distribution of 385 disease mutations in ExAC is shown in white. Violin plots show the density distribution of alleles segregating in these samples, whereas boxes indicate the proportion of sites for which the deleterious mutation was not observed. All distributions differ significantly from one another (i.e., all p-values are $< 10^{-15}$ by a Kolmogorov-Smirnov test).

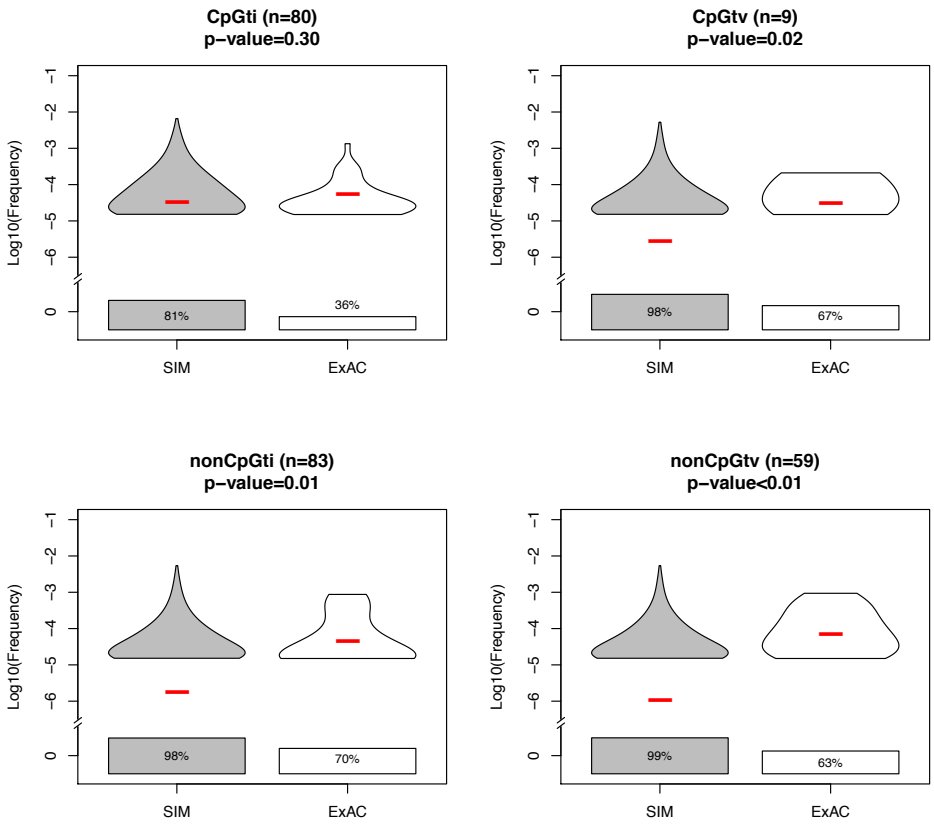


Fig S5. Expected and observed distributions of disease allele frequencies (excluding mutations in *CFTR* and *DHCR7*). The four panels correspond to four different mutation types. The title of the panel indicates the mutation type, followed by the total number of mutations, with p-value for the difference between observed and expected mean frequencies below it. Results in grey (SIM) were obtained from simulations considering a plausible demographic model for European populations [21] (see text). The observed values estimated from 33,370 individuals of European ancestry from ExAC are shown in white. As opposed to Fig 1, we did not include mutations present in two genes (*CFTR* and *DHCR7*) that were outliers in the gene-level analysis (Fig 3) and were reported elsewhere [20] to have healthy individuals found to be homozygous for a deleterious allele. As in Fig 1, violin plots show the density distribution of variable sites, while boxes indicate the proportion of sites for which the wild, non-deleterious mutation is fixed. Means (considering both segregating and fixed mutations) are indicated with red horizontal bars.

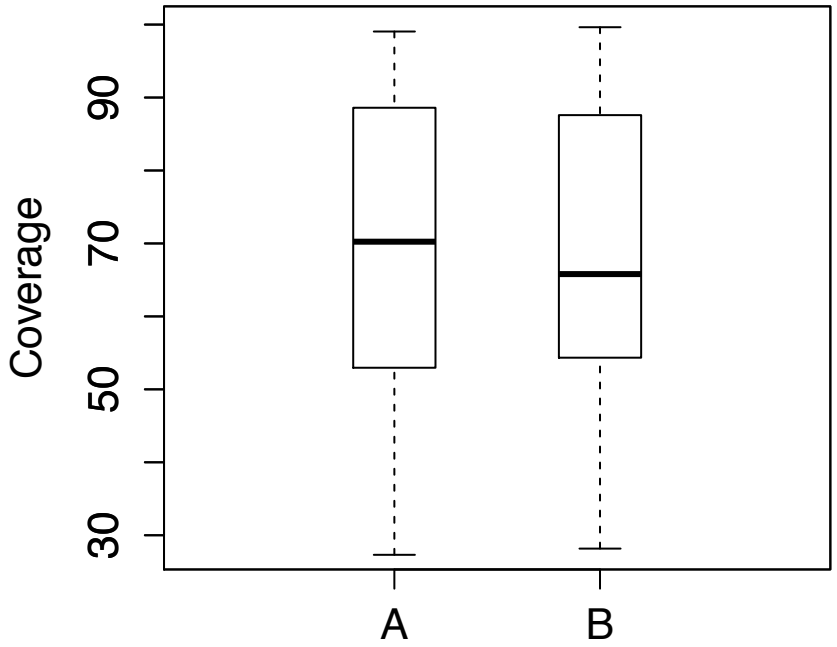


Fig S6. Depth of coverage for 385 mutations in ExAC known to cause lethal, Mendelian diseases. Box plots show the mean (black bar) and the lower and upper quartiles for (A) the 248 sites with non-zero sample frequencies in ExAC, for which the number of sequenced non-Finnish European individuals was reported (N = 32,881) and (B) for the 137 sites for which we did not have this information. Since distributions of depth of coverage are similar between datasets, we assumed that 32,881 individuals were sequenced on average at all sites, and used this number to subsample simulations to match the sample size of the ExAC data.