

1 The population genetics of human disease: the case of 2 recessive, lethal mutations

3 Carlos Eduardo G. Amorim^{1,2*}, Ziyue Gao³, Zachary Baker⁴, José Francisco Diesel⁵,
4 Yuval B. Simons¹, Imran S. Haque⁶, Joseph Pickrell^{1,7+}, Molly Przeworski^{1,4+}

5 ¹ Department of Biological Sciences, Columbia University, New York, NY

6 ² CAPES Foundation, Ministry of Education of Brazil, Brasília, DF, Brazil

7 ³ Howard Hughes Medical Institution, Stanford University, Stanford, CA

8 ⁴ Department of Systems Biology, Columbia University, New York, NY

9 ⁵ Universidade Federal de Santa Maria, Santa Maria, RS, Brazil

10 ⁶ Counsyl, 180 Kimball Way, South San Francisco, CA. Current address: Freenome, 201 Gateway
11 Blvd, South San Francisco, CA

12 ⁷ New York Genome Center, New York, NY

13 ⁺These authors co-supervised this work

14 ^{*} To whom correspondence should be addressed (guerraamorim@gmail.com); Current address:
15 Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY.

16

17

1 **Abstract**

2 Do the frequencies of disease mutations in human populations reflect a simple balance between
3 mutation and purifying selection? What other factors shape the prevalence of disease mutations?
4 To begin to answer these questions, we focused on one of the simplest cases: recessive mutations
5 that alone cause lethal diseases or complete sterility. To this end, we generated a hand-curated set
6 of 417 Mendelian mutations in 32 genes, reported to cause a recessive, lethal Mendelian disease.
7 We then considered analytic models of mutation-selection balance in infinite and finite populations
8 of constant sizes and simulations of purifying selection in a more realistic demographic setting, and
9 tested how well these models fit allele frequencies estimated from 33,370 individuals of European
10 ancestry. In doing so, we distinguished between CpG transitions, which occur at a substantially
11 elevated rate, and three other mutation types. The observed frequency for CpG transitions is
12 slightly higher than expectation but close, whereas the frequencies observed for the three other
13 mutation types are an order of magnitude higher than expected. This discrepancy is even larger
14 when subtle fitness effects in heterozygotes or lethal compound heterozygotes are taken into
15 account. In principle, higher than expected frequencies of disease mutations could be due to
16 widespread errors in reporting causal variants, compensation by other mutations, or balancing
17 selection. It is unclear why these factors would have a greater impact on variants with lower
18 mutation rates, however. We argue instead that the unexpectedly high frequency of disease
19 mutations and the relationship to the mutation rate likely reflect an ascertainment bias: of all the
20 mutations that cause recessive lethal diseases, those that by chance have reached higher
21 frequencies are more likely to have been identified and thus to have been included in this study.
22 Beyond the specific application, this study highlights the parameters likely to be important in
23 shaping the frequencies of Mendelian disease alleles.

1 **Author Summary**

2 What determines the frequencies of disease mutations in human populations? To begin to answer
3 this question, we focus on one of the simplest cases: mutations that cause completely recessive,
4 lethal Mendelian diseases. We first review theory about what to expect from mutation and
5 selection in a population of finite size and further generate predictions based on simulations using
6 a realistic demographic scenario of human evolution. For a highly mutable type of mutations, such
7 as transitions at CpG sites, we find that the predictions are close to the observed frequencies of
8 recessive lethal disease mutations. For less mutable types, however, predictions substantially
9 under-estimate the observed frequency. We discuss possible explanations for the discrepancy and
10 point to a complication that, to our knowledge, is not widely appreciated: that there exists
11 ascertainment bias in disease mutation discovery. Specifically, we suggest that alleles that have
12 been identified to date are likely the ones that by chance have reached higher frequencies and are
13 thus more likely to have been mapped. More generally, our study highlights the factors that
14 influence the frequencies of Mendelian disease alleles.

15

1 Introduction

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

8 In this regard, rare, Mendelian diseases, which are caused by single highly penetrant and
9 deleterious alleles, are perhaps most amenable to investigation. A simple model for the persistence
10 of mutations that lead to Mendelian diseases is that their frequencies should reflect an equilibrium
11 between their introduction by mutation and elimination by purifying selection, i.e., that they should
12 be found at “mutation-selection balance” (MSB) [4]. In finite populations, random drift leads to
13 stochastic changes in the frequency of any mutation, so demographic history, in addition to
14 mutation and natural selection, also plays an important role in shaping the frequency distribution
15 of deleterious mutations [3].

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

20 For a subset of disease alleles that are recessive, an alternative model for their persistence in
21 the population is that there is an advantage to carrying one copy but a disadvantage to carrying
22 two or none, such that the alleles persist due to overdominance, a form of balancing selection. Well
23 known examples include sickle cell anemia, thalassemia and G6PD deficiency in populations living

1 where malaria is endemic [14]. The importance of overdominance in maintaining the high
2 frequency of disease mutations is unknown beyond these specific cases.

3 Here, we tested hypotheses about the persistence of mutations that cause lethal, recessive,
4 Mendelian disorders. This case provides a good starting point, because a large number of
5 Mendelian disorders have been mapped (e.g., genes have already been associated with >66% of
6 Mendelian disease phenotypes; [15]). Moreover, while the fitness effects of most diseases are hard
7 to estimate, for recessive lethal diseases, the selection coefficient is clearly 1 for homozygote
8 carriers in the absence of modern medical care (which, when available, became so only in the last
9 couple of generations, a timescale that is much too short to substantially affect disease allele
10 frequencies). Moreover, assuming mutation-selection balance in an infinite population would
11 suggest that, given a per base pair (bp) mutation rate u on the order of 10^{-8} per generation [16], the
12 frequency of such alleles would be \sqrt{u} , i.e., $\sim 10^{-4}$ [4]. Thus, sample sizes in human genetics are now
13 sufficiently large that we should be able to observe recessive disease alleles segregating in
14 heterozygote carriers.

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

1 Results

2 Mendelian recessive disease allele set

3 We relied on two datasets, one that describes 173 autosomal recessive diseases [17] and
4 another from a genetic testing laboratory (Counsyl; <<https://www.counsyl.com/>>) that includes
5 110 recessive diseases of clinical interest. From these lists, we obtained a set of 44 “recessive
6 lethal” diseases associated with 45 genes (Table S1), requiring that at least one of the following
7 conditions is met: (i) in the absence of treatment, the affected individuals die of the disease before
8 reproductive age, (ii) reproduction is completely impaired in patients of both sexes, (iii) the
9 phenotype includes severe mental retardation that in practice precludes reproduction, or (iv) the
10 phenotype includes severely compromised physical development, again precluding reproduction.

11 Based on clinical genetics datasets and the medical literature (see Methods for details), we
12 were able to confirm that 417 Single Nucleotide Variants (SNVs) in 32 (of the 44) genes had been
13 reported with compelling evidence of association to the severe form of the corresponding disease
14 and an early-onset, as well as no indication of effects in heterozygote carriers (Table S2). By this
15 approach, we obtained a set of mutations for which, at least in principle, there is no heterozygote
16 effect, i.e., for which the dominance coefficient $h = 0$ in a model with relative fitness of 1 for the
17 homozygote for the reference allele, $1-hs$ for the heterozygote, and $1-s$ for the homozygote for the
18 deleterious allele, and the selective coefficient s is 1.

19 A large subset of these mutations (29.3%) consists of transitions at CpG sites (henceforth
20 CpGti), which occur at a highly elevated rates (~17-fold higher on average) compared to other
21 mutation types, namely CpG transversions, and non-CpG transitions and transversions [16]. This
22 proportion is in agreement with previous estimates for a smaller set of disease genes [18] and for
23 *DMD* [19].

1

2 **Empirical distribution of disease alleles in Europe**

3 Allele frequency data for the 417 variants were obtained from the Exome Aggregation
4 Consortium (ExAC) for 60,706 individuals, of whom 33,370 are non-Finnish Europeans [20]. Out of
5 the 417 variants associated with putative recessive lethal diseases, three were found homozygous
6 in at least one individual in this dataset (rs35269064, p.Arg108Leu in *ASS1*; rs28933375,
7 p.Asn252Ser in *PRF1*; and rs113857788, p.Gln1352His in *CFTR*). Available data quality information
8 for these variants does not suggest genotype calling artifacts (Table S2). Since these diseases have
9 severe symptoms that lead to early death without treatment and these ExAC individuals are
10 healthy (i.e., do not manifest severe Mendelian diseases) [20], the reported mutations are likely
11 errors in pathogenicity classification or cases of incomplete penetrance (see a similar observation
12 for *CFTR* and *DHCR7* in [21]). We therefore excluded them from our analyses. In addition to the
13 mutations present in homozygotes, we also filtered out sites that had lower coverage in ExAC (see
14 Methods), resulting in a final dataset of 385 variants in 32 genes (Table S2).

15 Genotypes for a subset (91) of these mutations were also available for a larger sample size
16 (76,314 individuals with self-reported European ancestry) generated by the company Counsyl
17 (Table S3). A comparison of the allele frequencies in this larger dataset to that of ExAC suggests
18 that the allele frequencies for individual variants are concordant between the two datasets
19 (Pearson's correlation coefficient of 0.79, Fig S1) and that the overall distributions do not differ
20 appreciably (Kolmogorov-Smirnov test, p-value = 0.23). Thus, both data sets appear to reflect the
21 general distribution of these disease alleles in Europeans. In what follows, we focus on ExAC, which
22 includes a greater number of disease mutations.

23

1 **Models of mutation-selection balance**

2 To generate expectations for the frequencies of these disease mutations under mutation-
3 selection balance, we considered models of infinite and finite populations of constant size [3], and
4 conducted forward simulations using a plausible demographic model for African and European
5 populations [22] (see Methods for details). In all these models, there is a wild-type allele (A) and a
6 deleterious allele (a, which could also represent a class of distinct deleterious alleles with the same
7 fitness effect) at each site, such that the relative fitness of individuals of genotypes AA, Aa, or aa is
8 given respectively by:

- 9 • $w_{AA}=1$;
- 10 • $w_{Aa}=1-hs$;
- 11 • $w_{aa}=1-s$;

12 The mutation rate from A to a is u ; we assume that there are no back mutations.

13 For a constant population of infinite size, Wright [23] showed that under these conditions,
14 there exists a stable equilibrium between mutation and selection, when the selection pressure is
15 sufficiently strong ($s \gg u$). In particular, when the deleterious effect of allele a is completely
16 recessive ($h=0$), its equilibrium frequency q is given by:

$$17 \quad q = \sqrt{u/s}. \quad (1)$$

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

$$20 \quad \bar{q} = \frac{\Gamma(2Nu + 1/2)}{\sqrt{2Ns}\Gamma(2Nu)}, \quad (2)$$

$$21 \quad \sigma_q^2 = u/s - \bar{q}^2, \quad (3)$$

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

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

$$6 \quad \bar{q} \approx \sqrt{u/s}, \quad (4)$$

7 which is independent of the population size and equal to the equilibrium frequency in an infinite
8 population, i.e., the right hand side of eq. (1). The important difference between the two models
9 above is that in a finite population, there is a distribution of frequencies q (because of genetic drift),
10 whose variance is given in eq. (3), rather than a single value, as in an infinite population.

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

$$13 \quad \bar{q} \approx u\sqrt{2\pi N/s}, \quad (5)$$

14 which depends on the population size [3].

15 We note that Nei [3] assumed a Wright-Fisher population, so there was no distinction between
16 the census and the effective population size. However, when the two differ, it is the effective
17 population size that governs the dynamics of deleterious alleles, so the N in the analytical results in
18 fact represents the effective population size. In humans, the mutation rate at each bp is very small
19 (on the order of 10^{-8} [16]) and the effective population size not that large, even recently [24,25], so
20 the second approximation should apply when considering each single site independently.

21 The expectation and variance of the frequency of a fatal, fully recessive allele (i.e., $s=1$, $h=0$) are
22 then given by:

$$\bar{q} = u\sqrt{2\pi N}, \quad (6)$$

2 and

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

4 This single site model implicitly ignored the existence of compound heterozygosity in modeling the
5 strength of selection acting on an individual site.

6

7 **Comparing mutation-selection balance models**

8 Although an infinite population size has often been assumed when modeling deleterious
9 allele frequencies (e.g. [5,26-29]), predictions under this assumption can differ markedly from
10 what is expected from models of finite population sizes, assuming plausible parameter values for
11 humans. For example, the long-term estimate of the effective population size from total
12 polymorphism levels is ~20,000 individuals (assuming a mutation rate of 1.2×10^{-8} per bp per
13 generation [16] and diversity levels of 0.1% [30]). In this case and considering a mutation rate of
14 1.5×10^{-8} for exons (which have a higher mutation rate than the rest of the genome, because of
15 their base composition [31]), the average deleterious allele frequency in the model of finite
16 population size is ~23-fold lower than that in the infinite population model (Fig 1).

17 Because the human population size has not been constant and changes in the population
18 size can affect the frequencies of deleterious alleles in the population (e.g. [2,32]), we also
19 simulated the population dynamics of disease alleles under a plausible demographic model for
20 European populations [22]. Assuming a mutation rate of 1.5×10^{-8} per bp, the mean allele
21 frequency of a lethal, recessive disease allele obtained from this model was 9.91×10^{-6} , ~1.86-fold
22 higher than expected for a constant population size model with $N = 20,000$ (Fig 1). The mean
23 frequency seen in simulations instead matches the expectation for a constant population size of

1 69,537 individuals (see Methods and Fig S2A). This finding is expected: the estimated effective
2 population size of 20,000 is based on total genetic diversity; assuming that most of this variation is
3 neutral, it therefore reflects an average timescale over millions of years. For recessive lethal
4 mutations, which are relatively rapidly purged by natural selection, a more recent time depth is
5 relevant (e.g., [1]). Indeed, in our simulations, most of the disease alleles (65.6%) segregating at
6 present arose very recently, such that they were not segregating in the population 205 generations
7 ago, the time point after which N_e is estimated to have increased from 9,300 to 512,000 individuals
8 [22].

9 Increasing the effective population size in a constant size model is not enough to capture the
10 dynamics of disease alleles appropriately, however. For example, if we compare simulation results
11 obtained under the more complex Tennesen et al. [22] demographic model to those for
12 simulations of a constant population size of $N = 69,537$, the mean allele frequencies match, but the
13 distributions of allele frequencies are significantly different (Kolmogorov-Smirnov test, p-value <
14 10^{-15} ; Fig S2B and S2C). These findings thus confirm the importance of incorporating demographic
15 history into models for understanding the population dynamics of disease alleles [5,33,34]. In what
16 follows, we therefore test the fit of the more realistic demographic model [22] (and variants of it)
17 to the observed allele frequencies.

18

19 **Comparing empirical and expected distributions of disease alleles**

20 The mutation rate from wild-type allele to disease allele, u , is a critical parameter in
21 predicting the frequencies of a deleterious allele [4,35]. To model disease alleles, we considered
22 four mutation types separately, with the goal of capturing most of the fine-scale heterogeneity in
23 mutation rates [24,36-38]: transitions in methylated CpG sites (CpGti) and three less mutable

1 types, namely transversions in CpG sites (CpGtv) and transitions and transversions outside a CpG
2 site (nonCpGti and nonCpGtv, respectively). In order to control for the methylation status of CpG
3 sites, we excluded 12 CpGti that occurred in CpG islands, which tend not to be methylated and thus
4 are likely to have a lower mutation rate [36] (following Moorjani et al. [39]). To allow for
5 heterogeneity in mutation rates within each one of these four classes considered, we modeled the
6 within-class variation in mutation rates according to a lognormal distribution (see details in
7 Methods and [24]).

8 For each mutation type, we then compared the mean allele frequency obtained from
9 simulations to what is observed in ExAC, running 100,000 replicates. To this end, we matched
10 simulations to the empirical data with regard to the number of individuals sampled and number of
11 mutations observed of each mutation type and focused the analysis on the largest sample of the
12 same common ancestry, namely Non-Finnish Europeans ($n = 33,370$) (Fig 2A). We find significant
13 differences between empirical and expected mean frequencies for nonCpGti (50-fold higher on
14 average; two-tailed p-value $< 1.4 \times 10^{-4}$; see Methods for details) and nonCpGtv (24-fold higher on
15 average, two-tailed p-value $< 1.2 \times 10^{-4}$), to a lesser extent for CpGtv (9-fold higher on average, two-
16 tailed p-value = 0.04). The mean frequency for CpGti is also somewhat higher than expected, but
17 insignificantly so (2-fold higher on average, two-tailed p-value = 0.18). Intriguingly, the
18 discrepancy between observed and expected frequencies becomes smaller as the mutation rate
19 increases (Fig 2B).

20 Two additional factors that we have not included in our model should further decrease the
21 predicted frequencies of disease alleles. Given that frequencies in ExAC are already unexpectedly
22 high, these factors would only exacerbate the discrepancy between observed and expected
23 frequencies of deleterious alleles. First, we have ignored the effects of compound heterozygosity,

1 the case in which combinations of two distinct pathogenic alleles in the same gene lead to lethality.
2 This phenomenon is known to be common [40], and indeed 58.4% of the 417 disease mutations
3 considered in this study were initially identified in compound heterozygotes. In the presence of
4 compound heterozygosity, each deleterious mutation will be selected against not only when
5 present in two copies within the same individual, but also in the presence of lethal mutations at
6 other sites of the same gene. Since the purging effect of selection against compound heterozygotes
7 was not modeled in simulations, we would predict the frequency of a deleterious mutation to be
8 even lower than shown (e.g., in Fig 2A).

9 In order to model the effect of compound heterozygosity in our simulations, we re-ran our
10 simulations, but focusing on a gene rather than a single site and so considering the sum of
11 frequencies of all known recessive lethal alleles within a gene. In these simulations, we used the
12 same set-up as in the site level analysis, except for the mutation rate, U , which is now the sum of the
13 mutation rates u_j at each site j that is known to cause a severe and early onset form of the disease
14 (Table S2; see Methods for details). This approach does not consider the contribution of other
15 mutations in the genes that cause the mild and/or late onset forms of the disease, and implicitly
16 assumes that all combinations of known recessive lethal alleles of the same gene have the same
17 fitness effect as homozygotes. Comparing observed frequencies of disease alleles for each gene to
18 predictions generated by simulation, one third of the 27 genes differ from the expected distribution
19 at the 5% level, with a clear overall trend for observed frequencies to be above expectation (Table
20 S4; Fig 3; Fisher's combined probability test p-value $< 10^{-14}$).

21 This finding is even more surprising than it may seem, because we are far from knowing the
22 complete mutation target for each gene, i.e., all the sites at which mutations could cause the disease.
23 If there are additional, undiscovered sites in the gene at which mutations are fatal when carried in

1 combination with a known recessive lethal mutation, the purging effect of purifying selection on
2 the known mutations will be under-estimated in our simulations leading us to over-estimate the
3 expected frequencies of the known mutations in simulations. Therefore, our predictions are, if
4 anything, an over-estimate of the expected allele frequency, and the discrepancy between predicted
5 and the observed is likely even larger than what we found.

6 The other factor that we did not consider in simulations but would reduce the expected
7 allele frequencies is a subtle fitness decrease in heterozygotes. To evaluate potential fitness effects
8 in heterozygotes when none had been documented in humans, we considered the phenotypic
9 consequences of orthologous gene knockouts in mouse. We were able to retrieve information on
10 phenotypes for both homozygote and heterozygote mice for only eight out of the 32 genes, namely
11 *ASS1*, *CFTR*, *DHCR7*, *NPC1*, *POLG*, *PRF1*, *SLC22A5*, and *SMPD1*. For all eight, homozygote knockout
12 mice presented similar phenotypes as affected humans, and heterozygotes showed a milder but
13 detectable phenotype (Table S5). The magnitude of the heterozygote effect of these mutations in
14 humans is unclear, but the finding with knockout mice makes it plausible that there exists a very
15 small fitness decrease in heterozygotes in humans as well, potentially not enough to have been
16 recognized in clinical investigations but enough to have a marked impact on the allele frequencies
17 of the disease mutations. Indeed, even if the fitness effect in heterozygotes were as small as $h = 1\%$,
18 a 61% decrease in the mean allele frequency of the disease allele is expected relative to the case
19 with complete recessivity ($h = 0$) (Fig S3).

20 **Discussion**

21 To investigate the population genetics of human disease, we focused on mutations that cause
22 Mendelian, recessive disorders that lead to early death or completely impaired reproduction. We

1 sought to understand to what extent the frequencies of these mutations fit the expectation based
2 on a simple balance between the input of mutations and the purging by purifying selection, as well
3 as how other mechanisms might affect these frequencies. Many studies implicitly or explicitly
4 compare known disease allele frequencies to expectations from mutation-selection balance [5,26-
5 29]. In this study, we tested whether known recessive lethal disease alleles as a class fit these
6 expectations, and found that, under a sensible demographic model for European population history
7 with purifying selection only in homozygotes, the expectations fit the observed disease allele
8 frequencies poorly: the mean empirical frequencies of disease alleles are substantially above
9 expectation for all mutation types (although not significantly so for CpGti), and the fold increase in
10 observed mean allele frequency in relation to the expectation decreases with increased mutation
11 rate (Fig 2). Furthermore, including possible effects of compound heterozygosity and subtle fitness
12 decrease in heterozygotes will only exacerbate the discrepancy.

13 In principle, higher than expected disease allele frequencies could be explained by at least six
14 (non-mutually exclusive) factors: (i) widespread errors in reporting the causal variants; (ii)
15 misspecification of the demographic model, (iii) misspecification of the mutation rate; (iv)
16 reproductive compensation; (v) overdominance of disease alleles; and (vi) low penetrance of
17 disease mutations. Because widespread mis-annotation of the causal variants in disease mutation
18 databases has previously been reported [20,41,42], we tried to minimize the effect of such errors
19 on our analyses by filtering out any case that lacked compelling evidence of association with a
20 recessive lethal disease, reducing our initial set of 769 mutations to 385 in which we had greater
21 confidence (see Methods for details).

22 We also explored the effects of mis-specifying recent demographic history or the mutation rate.
23 Based on very large samples, it has been estimated that population growth in Europe could have

1 been stronger than what we considered in our simulations [43,44]. When we consider higher
2 growth rates, such that the current effective population size is up to 20-fold larger than that of the
3 original model, we observe an increase in the expected frequency of recessive disease alleles and a
4 larger number of segregating sites (Fig S4, columns A-F). However, the impact of larger growth rate
5 is insufficient to explain the observed discrepancy: the allele frequencies observed in ExAC are still
6 on average an order of magnitude larger than expected based on a model with a 20-fold larger
7 current effective population size than the one initially considered [22] (Fig S4). Similarly,
8 population substructure within Europe would only increase the number of homozygotes relative to
9 what was modeled in our simulations (through the Wahlund effect [45]) and expose more
10 recessive alleles to selection, thus decreasing the expected allele frequencies and exacerbating the
11 discrepancy that we report.

12 In turn, to explore the effects of error in the mutation rate, we considered a 50% higher mean
13 mutation rate than what has been estimated for exons [31], beyond what seems plausible based on
14 current estimates on human mutation rates [39]. Except for the mean mutation rate (now set to
15 2.25×10^{-8}), all other parameters used for these simulations (i.e. the variance in mutation rate
16 across simulations, the demographic model [22], absence of selective effect in heterozygotes, and
17 selection coefficient) were kept the same as the ones used for generating Fig S4 column A. The
18 observed mean frequency remains significantly above what we predict using this high mutation
19 rate and qualitative conclusions are unchanged (Fig S4, column G).

20 Another factor to consider is that for disease phenotypes that are lethal very early on in life,
21 there may be partial or complete reproductive compensation (e.g. [46]). This phenomenon would
22 decrease the fitness effects of the recessive lethal mutations and could therefore lead to an increase

1 in the allele frequency in data relative to what we predict for a selection coefficient of 1. There are
2 no reasons, however, for this phenomenon to correlate with the mutation rate, as seen in Fig 2B.

3 The other two factors, overdominance and low penetrance, are likely explanations for a subset
4 of cases. For instance, *CFTR*, the gene in which mutations lead to cystic fibrosis, is the furthest
5 above expectation (p-value = 0.002; Fig 3). It was long noted that there is an unusually high
6 frequency of the *CFTR* deletion $\Delta F508$ in Europeans, which led to speculation that disease alleles of
7 this gene may be subject to over-dominance ([47,48], but see [49]). Regardless, it is known that
8 disease mutations in this gene can complement one another [8,9] and that modifier loci in other
9 genes also influence their penetrance [9,12]. Consistent with variable penetrance, Chen et al. [21]
10 identified three purportedly healthy individuals carrying two copies of disease mutations in this
11 gene. Similarly, *DHCR7*, the gene associated with the Smith-Lemli-Opitz syndrome, is somewhat
12 above expectation in our analysis (p-value = 0.052; Fig 3) and healthy individuals were found to be
13 homozygous carriers of putatively lethal disease alleles in other studies [21]. These observations
14 make it plausible that, in a subset of cases (particularly for *CFTR*), the high frequency of deleterious
15 mutations associated with recessive, lethal diseases are due to genetic interactions that modify the
16 penetrance of certain recessive disease mutations. It is hard to assess the importance of this
17 phenomenon in driving the general pattern that we observe, but two factors argue against it being
18 a sufficient explanation for what we find at the level of single sites. First, when we remove 130
19 mutations in *CFTR* and 12 in *DHCR7*, the two genes that were outliers at the gene-level (Fig 3; Table
20 S4) and for which we have evidence for incomplete penetrance [21], the discrepancy between
21 observed and expected allele frequencies is barely impacted (Fig S5). Moreover, there is no obvious
22 reason why the degree of incomplete penetrance would vary systematically with the mutation rate
23 of a site, as we observe (Fig 2B).

1 Instead, it seems plausible that there is an ascertainment bias in disease allele discovery and
2 mutation identification [48,50,51]. Unlike mis-sense or protein-truncating variants, Mendelian
3 disease mutations cannot be readily annotated based solely on DNA sequences, and their
4 identification requires reliable diagnosis of affected individuals (usually in more than one
5 pedigree) followed by rigorous mapping of the underlying gene/mutation. Therefore, those
6 mutations that have been identified to date are likely the ones that are segregating at higher
7 frequencies in the population. Indeed, the guidelines for the interpretation of the pathogenicity of
8 sequence variants by the American College of Medical Genetics and Genomics [52] may bias the
9 identification of disease alleles towards high frequency ones. Moreover, mutation-selection balance
10 models predict that the frequency of a deleterious mutation should correlate with the mutation
11 rate. Together, these considerations suggest that disease variants of a highly mutable class, such as
12 CpGti, are more likely to have been mapped and that the mean frequency of mapped mutations
13 should be slightly above but close to that of all disease mutations in that class. We would further
14 predict that, in contrast, less mutable disease mutations are less likely to have been discovered to
15 date, and that the mean frequency of the subset of mutations that have been identified will far
16 exceed that of all mutations in that class.

17 To quantify these effects, we modeled the ascertainment of disease mutations both analytically
18 and in simulations (see Methods). As expected, we find that for a given mutation type, the average
19 allele frequency of mutations that have been identified is always higher than that of all existing
20 mutations (Table 1). Furthermore, comparison across different mutation types reveals that a
21 higher mutation rate increases the probability of disease mutations being ascertained (Table 1 and
22 Fig S6) and decreases the magnitude of bias in estimated allele frequency relative to the mutation
23 class as a whole (Table 1). In summary, among all the possible aforementioned explanations for the

1 observed discrepancy between empirical and expected mean allele frequencies, the ascertainment
2 bias hypothesis is the only one that also explains why it is more pronounced for less mutable
3 mutation types (Fig 2B).

4 One clear implication of this hypothesis is that there are numerous sites at which mutations
5 cause recessive lethal diseases yet to be discovered, particular at non-CpG sites. More generally,
6 this ascertainment bias complicates the interpretation of observed allele frequencies in terms of
7 the selection pressures acting on disease alleles. Beyond this specific point, our study illustrates
8 how the large sample sizes now made available to researchers in the context of projects like ExAC
9 [20] can be used not only for direct discovery of disease variants, but also to test why disease
10 alleles are segregating in the population and to understand at what frequencies we might expect to
11 find them.

12 **Methods**

13 **Disease allele set**

14 In order to identify single nucleotide variants within the 42 genes associated with lethal,
15 recessive Mendelian diseases (Table S1), we initially relied on the ClinVar dataset [53] (accessed on
16 June 3rd, 2015). We filtered out any variant that is an indel or a more complex copy number variant
17 or that is ever classified as benign or likely benign in ClinVar (whether or not it is also classified as
18 pathogenic or likely pathogenic). By this approach, we obtained 769 SNVs described as pathogenic
19 or likely pathogenic. For each one of these variants, we searched the literature for evidence that it
20 is exclusively associated to the lethal and early onset form of the disease and was never reported as
21 causing the mild and/or late-onset form of the disease. We considered effects in the absence of
22 medical treatment, as we were interested in the selection pressures acting on the alleles over

1 evolutionary time scales rather than in the last one or two generations, i.e., the period over which
2 treatment became available for some of diseases considered. To evaluate the impact of treatment,
3 we decreased s from 1 to 0 (i.e., we assumed a complete absence of selective effects due to
4 treatment) in the last three generations and compared the mean allele frequencies across 100,000
5 simulations implemented with or without this readjustment in selection coefficient. Because of the
6 stochastic nature of the simulations, we repeated this pairwise comparison 10 times in order to get
7 a range of expected increase in allele frequencies. We observe only a minor increase in the mean
8 allele frequency (5.5% at most) across the 10 replicates. This simulation procedure corresponds to
9 a scenario in which there is an extremely effective treatment for all diseases for the past three
10 generations, which is a vast overestimate of the effect and length of treatment for the disease set
11 that we consider.

12 Variants with mention of incomplete penetrance (i.e. for which homozygotes were not
13 always affected) or with known effects in heterozygote carriers were removed from the analysis.
14 This process yielded 417 SNVs in 32 genes associated with distinct Mendelian recessive lethal
15 disorders (Table S2). Although these mutations were purportedly associated with complete
16 recessive diseases, we sought to examine whether there would be possible, unreported effects in
17 heterozygous carriers. To this end, we used the Mouse Genome Database (MGD) [54] (accessed July
18 29th, 2015) and were able to retrieve information for both homozygote and heterozygote mice for
19 eight out of the 32 genes (all of which with a homologue in mice) (Table S5).

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

1

2 **Genetic datasets**

3 The Exome Aggregation Consortium (ExAC) [20] was accessed on August 9th, 2016. The
4 data consist of genotype frequencies for 60,706 individuals, assigned after Principal Component
5 Analysis to one of seven population labels: African (n=5,203), East Asian (n=4,327), Finnish
6 (n=3,307, Latino (n=5,789), Non-Finnish European (n=33,370), South Asian (n=8,256) and “other”
7 (n=454) [20]. We focused our analyses on those individuals of Non-Finnish European descent,
8 because they constitute the largest sample size from a single ancestry group. We note that, some
9 disease mutations, for instance, those in *ASPA*, *HEXA* and *SMPD1*, are known to be especially
10 prevalent in Ashkenazi Jewish populations, which could potentially bias our results if Ashkenazi
11 Jewish individuals constituted a great portion of the sample we considered. However, this sample
12 includes only ~2,000 (~6%) Ashkenazi individuals (Dr. Daniel MacArthur, personal
13 communication).

14 From the initial 417 mutations, we filtered out three that were homozygous in at least one
15 individual in ExAC and 29 that had lower coverage, i.e., fewer than 80% of the individuals were
16 sequenced to at least 15x. This approach left us with a set of 385 mutations with a minimum
17 coverage of 27x per sample and an average coverage of 69x per sample (Table S2). For 248 sites
18 with non-zero sample frequencies, ExAC reported the number of non-Finnish European individuals
19 that were sequenced, which was on average 32,881 individuals [20]. For the remaining 137 sites,
20 we did not have this information. Nonetheless, the mean coverage across all samples is reported for
21 each site and does not differ between the two sets of sites (Fig S7). We therefore assumed that
22 mean number of individuals covered for all sites was 32,881 [55] and used this number to obtain
23 sample frequencies from simulations, as explained below.

1 A second genetic dataset was obtained from Counsyl (<<https://www.counsyl.com/>>).
2 Counsyl is a commercial genetic screening laboratory that offers, among other products, the
3 “Family Prep Screen”, a genetic screening test intended to detect carrier status for up to 110
4 recessive Mendelian diseases in couples that are planning to have a child. A subset of 294,000 of its
5 customers was surveyed by genotyping or sequencing for “routine carrier screening”. This subset
6 excludes individuals with indications for testing because of known personal or family history of
7 Mendelian diseases, infertility, and consanguinity. It therefore represents a more random (with
8 regard to the presence of disease alleles), population-based survey. For these individuals, we had
9 details on self-reported ancestry (14 distinct ethnic/ancestry/geographic groups) and the allele
10 frequencies for 98 mutations that match those that passed our variant selection criteria described
11 above, of which 91 are also sequenced to high coverage in the ExAC database (Table S2). We
12 focused our analysis of this dataset on the 76,314 individuals with self-reported Northern or
13 Southern European ancestry.

14

15 **Simulating the evolution of disease alleles with population size change**

16 We modeled the frequency of a deleterious allele in human populations by forward
17 simulations based on a crude but plausible demographic model for human populations from Africa
18 and Europe, inferred from exome data for African-Americans and European-Americans [22]. To
19 this end, we used a program described in [1]. In brief, the demographic scenario consists of an Out-
20 of-Africa demographic model, with changes in population size throughout the population history,
21 including a severe bottleneck in Europeans following the split from the African population and a
22 rapid, recent population growth in both populations [22]. As in Simons et al. [1], we simulated

1 genetic drift and two-way gene flow between Africans and Europeans in recent history. Negative
2 selection acting on a single bi-allelic site was modeled as in the analytic models.

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

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

14 When implementing the model, we considered a mean mutation rate u of 1.5×10^{-8} per bp,
15 per generation, as has been estimated for exons [31], as well as mutation rates for four distinct
16 mutation types (CpGti = 1.12×10^{-7} ; CpGtv = 9.59×10^{-9} ; nonCpGti = 6.18×10^{-9} ; and nonCpGtv =
17 3.76×10^{-9}) estimated from a large human pedigree study [16]. While these four categories capture
18 much of the variation in germline mutation rates across sites, a number of other factors (e.g., the
19 larger sequence context or the replication timing) also influence mutation rates, introducing
20 heterogeneity in the mutation rate within each class considered [24,36,37,56]. To allow for this
21 heterogeneity as well as for uncertainty in the point mutation rates estimates, in each simulation,
22 instead of using a fixed rate u for each mutation type, we drew the mutation rate M from a
23 lognormal distribution with the following parameters:

1
$$\log_{10} M | u \sim N(\log_{10} u - \frac{\sigma^2}{2} \ln(10), \sigma^2) \quad (8)$$

2 such that that $E[M]=u$. σ was set to 0.57 (following [24]).

3 For each mutation type, we then proceeded as follows:

4 (1) We ran two million simulations, thus obtaining the distribution of deleterious
5 allele frequencies expected for the European population.

6 (2) We sampled K allele frequencies from the two million simulations implemented
7 for each mutation type, where K is the number of identified mutations of that
8 type. Sample allele frequencies were simulated from these population
9 frequencies by Poisson sampling, so to match ExAC's number of chromosomes.

10 (3) We repeated step (2) 100,000 times, thus obtaining a distribution for the mean
11 allele frequency across K mutations.

12 To assess the significance of the deviation between observed and expected mean, we obtained a
13 two-tailed p-value, defined as $2 \times (r+1)/(100000+1)$, where r is the number of simulated allele
14 frequencies that were greater or equal to that of the empirical mean [57], for each mutation type
15 separately.

16 A well-known source of heterogeneity in mutation rate within the CpGti class is methylation
17 status, with a high transition rate seen only at methylated CpGs [18]. In our analyses, we tried to
18 control for the methylation status of CpG sites by excluding sites located in CpG islands (CGIs),
19 which tend to not be methylated [39]. The CGI annotation for hg19 was obtained from UCSC
20 Genome Browser (track "Unmasked CpG";
21 <<http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/cpgIslandExtUnmasked.txt.gz>>,
22 accessed in June 6th, 2016). BEDTools [58] was used to exclude those CpG sites located in CGIs. We

1 note that the CpGti estimate from [16] includes CGIs, and in that sense the average mutation rate
2 that we are using for CpGti may be a very slight underestimate of the mean rate for transitions at
3 methylated CpG sites.

4 Unless otherwise noted, the expectation assumes fully recessive, lethal alleles with complete
5 penetrance. Notably, by calculating the expected frequency one site at a time, we are ignoring
6 possible interaction between genes (i.e., effects of the genetic background) and among different
7 mutations within a gene (i.e., compound heterozygotes). These assumptions are relaxed in two
8 ways. In one analysis (Fig S3), we consider a very low selective effect in heterozygous individuals
9 ($h = 1\%$), reasoning that such an effect could plausibly go undetected in medical examinations and
10 yet would nonetheless impact the frequency of the disease allele. Second, when considering the
11 gene-level analysis (Fig 3), we implicitly allow for compound heterozygosity between any pair of
12 known lethal mutations. For this analysis, we ran 1000 simulations for a total mutation rate U per
13 gene that was calculated accounting for the heterogeneity and uncertainty in the mutation rates
14 estimates as follows: (i) For j sites in a gene known to cause a recessive lethal disease and that
15 passed our filtering criteria (Table S2), we drew a mutation rate u_j from the lognormal distribution,
16 as described above; (ii) We then took the sum of u_j as the total mutation rate U ; (iii) We then ran
17 one replicate with U as the mutation parameter, and other parameters as specified for site level
18 analysis. Because the mutational target size considered in simulations is only comprised of those
19 sites at which mutations are known to cause a lethal recessive disease, it is almost certainly an
20 underestimate of the true mutation rate—potentially by a lot. We note further that by this
21 approach, we are assuming that compound heterozygotes formed by any two lethal alleles have
22 fitness zero, i.e., that they are identical in their effects to homozygotes for any of the lethal alleles.
23 Moreover, we are implicitly ignoring the possibility of complementation, which is (somewhat)

1 justified by our focus on mutations with severe effects and complete penetrance (but see
2 Discussion). Since we were interested in understanding the effect of compound heterozygosity, for
3 this analysis, we did not consider the five genes in which only one mutation passed our filters
4 (*BCS1L*, *FKTN*, *LAMA3*, *PLA3G6*, and *TCIRG1*).

6 **Modeling the effect of the ascertainment bias in disease discovery**

7 To calculate the probability of ascertaining a recessive, lethal mutation, we assumed that all
8 currently known disease mutations were identified in a putative ascertainment study of sample
9 size n_a in a population with an inbreeding coefficient of F_a . Under this model, we can estimate P_a ,
10 the probability of ascertaining a disease mutation, as following:

11 For a disease allele (denoted as a) at frequency p in the present population, if we randomly
12 sample an individual with inbreeding coefficient of F , the probabilities of the three genotypes are:

$$13 \quad P(AA) = F((1 - p)) + (1 - F)(1 - p)^2, \quad (9)$$

$$14 \quad P(Aa) = (1 - F)2p(1 - p), \quad (10)$$

$$15 \quad P(aa) = Fp + (1 - F)p^2. \quad (11)$$

16 Thus, if n_a unrelated individual are surveyed, the probability of not seeing any homozygote
17 for the deleterious allele (which is the same as the probability of not being ascertained in this set)
18 is:

$$19 \quad P_{na} = (1 - p(aa))^{n_a}. \quad (12)$$

20 Therefore, the probability of ascertainment is

$$\begin{aligned} P_a &= 1 - P_{na} = 1 - (1 - p(aa))^{n_a} \\ &= 1 - [1 - Fp - (1 - F)p^2]^{n_a} \end{aligned}$$

1
$$= 1 - [(1 - p)(1 + p - Fp)]^{n_a}, \tag{13}$$

2 which is an increasing function with regard to p (the population allele frequency), F (the inbreeding
3 coefficient of the population under study) as well as n_a (the sample size of the putative
4 ascertainment study) (Fig S6).

5 We also demonstrate the relationship between the probability of ascertainment and
6 mutation rate using simulations of ascertainment bias implemented according to the following
7 steps:

- 8 1) For each of the four mutation types considered, we generated 10^6 allele frequencies q
9 from the results of the simulations based on a realistic demographic model [22].
- 10 2) We generated n_a independent diploid genotypes, given the allele frequencies from step 1
11 and an inbreeding coefficient F_a . We ran this step for a range of n_a and F_a values (Table
12 1).
- 13 3) With a given combination of n_a and F_a values, we identified the cases (out of the 10^6
14 observations from step 1) where at least one homozygote individual was observed in
15 step 2. These cases correspond to disease mutations that were ascertained; the
16 reasoning being that given complete penetrance, a recessive disease mutation can only
17 be identified if there is at least one affected individual in the studied population. With
18 this step, we calculated the probability of ascertainment by taking the fraction of cases
19 that satisfy the criteria above.
- 20 4) Finally, for each one of the 10^6 simulations from step 1, we generated a sample allele
21 frequency of the disease mutation, matching ExAC's sample size (i.e., considering $2n =$
22 65,762 chromosomes). We can then compare q_u , the unbiased average allele frequency

1 of all disease mutations, to q_a , the mean frequency of the subset of cases ascertained in
2 step 3, i.e., those cases for which at least one homozygote individual is observed.

3 These simulations were meant to illustrate the likely impact of ascertainment bias, rather
4 than to precisely describe the disease mutation identification process or to quantify the expected
5 effect. Notably, we performed these simulations for single sites, so the criteria for ascertainment in
6 step 3 did not include the possibility of compound heterozygotes, despite the fact that most
7 (58.4%) of the 417 disease mutations included in our study were initially identified in compound
8 heterozygotes. However, this simulation framework could readily be extended in this direction and
9 it would not change our qualitative conclusion.

10 **Acknowledgements**

11 We thank Daniel MacArthur for his help with the ExAC data, Ellen Leffler for providing her
12 Python script (available at <https://github.com/cegamorim/PopGenHumDisease>), as well as
13 Aravinda Chakravarti, Damien Labuda and four anonymous reviewers for comments on an earlier
14 version of the manuscript. All codes and data to generate the figures in R [59] and the script used to
15 get the sequence context of each mutation are available at
16 <https://github.com/cegamorim/PopGenHumDisease>. The code to run the simulations is available
17 at <https://github.com/sellalab/ForwardSimulator>. Allele frequencies and other information for the
18 disease mutations employed in the analyses are in Tables S2 and S3.

19

1 Author Contributions

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

5 References

- 6 1. Simons YB, Turchin MC, Pritchard JK, Sella G (2014) The deleterious mutation load is insensitive
7 to recent population history. *Nat Genet* 46: 220-224.
- 8 2. Simons YB, Sella G (2016) The impact of recent population history on the deleterious mutation
9 load in humans and close evolutionary relatives. *Curr Opin Genet Dev* 41: 150-158.
- 10 3. Nei M (1968) The frequency distribution of lethal chromosomes in finite populations. *Proc Natl*
11 *Acad Sci U S A* 60: 517-524.
- 12 4. Gillespie JH (2004) *Population Genetics: A Concise Guide*. Baltimore, MD: Johns Hopkins
13 University Press.
- 14 5. Brandvain Y, Wright SI (2016) The Limits of Natural Selection in a Nonequilibrium World.
15 *Trends Genet* 32: 201-210.
- 16 6. Balick DJ, Do R, Cassa CA, Reich D, Sunyaev SR (2015) Dominance of Deleterious Alleles Controls
17 the Response to a Population Bottleneck. *PLoS Genet* 11: e1005436.
- 18 7. Beauchamp KA, Muzzey D, Wong KK, Hogan GJ, Karimi K, et al. (2016) Systematic Design and
19 Comparison of Expanded Carrier Screening Panels. *bioRxiv*.
- 20 8. Cormet-Boyaka E, Jablonsky M, Naren AP, Jackson PL, Muccio DD, et al. (2004) Rescuing cystic
21 fibrosis transmembrane conductance regulator (CFTR)-processing mutants by
22 transcomplementation. *Proc Natl Acad Sci U S A* 101: 8221-8226.
- 23 9. Rapino D, Sabirzhanova I, Lopes-Pacheco M, Grover R, Guggino WB, et al. (2015) Rescue of NBD2
24 mutants N1303K and S1235R of CFTR by small-molecule correctors and
25 transcomplementation. *PLoS One* 10: e0119796.
- 26 10. Andressoo JO, Jans J, de Wit J, Coin F, Hoogstraten D, et al. (2006) Rescue of progeria in
27 trichothiodystrophy by homozygous lethal Xpd alleles. *PLoS Biol* 4: e322.
- 28 11. Gallati S (2014) Disease-modifying genes and monogenic disorders: experience in cystic
29 fibrosis. *Appl Clin Genet* 7: 133-146.
- 30 12. Corvol H, Blackman SM, Boelle PY, Gallins PJ, Pace RG, et al. (2015) Genome-wide association
31 meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nat*
32 *Commun* 6: 8382.
- 33 13. Habara A, Steinberg MH (2016) Minireview: Genetic basis of heterogeneity and severity in
34 sickle cell disease. *Exp Biol Med (Maywood)* 241: 689-696.
- 35 14. Hedrick PW (2011) Population genetics of malaria resistance in humans. *Heredity (Edinb)* 107:
36 283-304.

- 1 15. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, et al. (2015) The Genetic Basis of
2 Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet* 97:
3 199-215.
- 4 16. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, et al. (2012) Rate of de novo mutations
5 and the importance of father's age to disease risk. *Nature* 488: 471-475.
- 6 17. Turner TN, Douville C, Kim D, Stenson PD, Cooper DN, et al. (2015) Proteins linked to autosomal
7 dominant and autosomal recessive disorders harbor characteristic rare missense mutation
8 distribution patterns. *Hum Mol Genet* 24: 5995-6002.
- 9 18. Cooper DN, Yousoufian H (1988) The CpG dinucleotide and human genetic disease. *Hum Genet*
10 78: 151-155.
- 11 19. Akalin N, Zietkiewicz E, Makalowski W, Labuda D (1994) Are CpG sites mutation hot spots in
12 the dystrophin gene? *Hum Mol Genet* 3.
- 13 20. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, et al. (2016) Analysis of protein-coding
14 genetic variation in 60,706 humans. *Nature* 536: 285-291.
- 15 21. Chen R, Shi L, Hakenberg J, Naughton B, Sklar P, et al. (2016) Analysis of 589,306 genomes
16 identifies individuals resilient to severe Mendelian childhood diseases. *Nat Biotechnol* 34:
17 531-538.
- 18 22. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, et al. (2012) Evolution and functional
19 impact of rare coding variation from deep sequencing of human exomes. *Science* 337: 64-69.
- 20 23. Wright S (1937) The Distribution of Gene Frequencies in Populations. *Proc Natl Acad Sci U S A*
21 23: 307-320.
- 22 24. Harpak A, Bhaskar A, Pritchard JK (2016) Mutation Rate Variation is a Primary Determinant of
23 the Distribution of Allele Frequencies in Humans. *PLoS Genet* 12: e1006489.
- 24 25. Browning Sharon R, Browning Brian L (2015) Accurate Non-parametric Estimation of Recent
25 Effective Population Size from Segments of Identity by Descent. *The American Journal of*
26 *Human Genetics* 97: 404-418.
- 27 26. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. *Trends Genet* 17: 502-
28 510.
- 29 27. Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-
30 common variant...or not? *Hum Mol Genet* 11: 2417-2423.
- 31 28. Zwick ME, Cutler DJ, Chakravarti A (2000) Patterns of genetic variation in Mendelian and
32 complex traits. *Annu Rev Genomics Hum Genet* 1: 387-407.
- 33 29. Cassa CA, Weghorn D, Balick DJ, Jordan DM, Nusinow D, et al. (2017) Estimating the Selective
34 Effect of Heterozygous Protein Truncating Variants from Human Exome Data. *Nat Genet* 49:
35 806-810.
- 36 30. The 1000 Genomes Consortium Project (2015) A global reference for human genetic variation.
37 *Nature* 526: 68-74.
- 38 31. Neale BM, Kou Y, Liu L, Ma/'ayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de
39 novo mutations in autism spectrum disorders. *Nature* 485: 242-245.
- 40 32. Fu W, O'Connor T, Jun G, Kang H, Abecasis G, et al. (2013) Analysis of 6,515 exomes reveals the
41 recent origin of most human protein-coding variants. *Nature* 493: 216-220.
- 42 33. Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, et al. (2008) Proportionally
43 more deleterious genetic variation in European than in African populations. *Nature* 451:
44 994-997.
- 45 34. Peischl S, Dupanloup I, Bosshard L, Excoffier L (2016) Genetic surfing in human populations:
46 from genes to genomes. *Curr Opin Genet Dev* 41: 53-61.

- 1 35. Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-
2 common variant...or not? *Hum Mol Genet* 11:2417-23.
- 3 36. Segurel L, Wyman MJ, Przeworski M (2014) Determinants of mutation rate variation in the
4 human germline. *Annu Rev Genomics Hum Genet* 15: 47-70.
- 5 37. Aggarwala V, Voight BF (2016) An expanded sequence context model broadly explains
6 variability in polymorphism levels across the human genome. *Nat Genet* 48: 349-355.
- 7 38. Hodgkinson A, Eyre-Walker A Variation in the mutation rate across mammalian genomes. *Nat*
8 *Rev Genet* 12: 756-766.
- 9 39. Moorjani P, Amorim CE, Arndt PF, Przeworski M (2016) Variation in the molecular clock of
10 primates. *Proc Natl Acad Sci U S A* 113: 10607-10612.
- 11 40. Kamphans T, Sabri P, Zhu N, Heinrich V, Mundlos S, et al. (2013) Filtering for compound
12 heterozygous sequence variants in non-consanguineous pedigrees. *PLoS One* 8: e70151.
- 13 41. Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, et al. (2012) Deleterious- and disease-allele prevalence
14 in healthy individuals: insights from current predictions, mutation databases, and
15 population-scale resequencing. *Am J Hum Genet* 91: 1022-1032.
- 16 42. Piton A, Redin C, Mandel JL (2013) XLID-causing mutations and associated genes challenged in
17 light of data from large-scale human exome sequencing. *Am J Hum Genet* 93: 368-383.
- 18 43. Gazave E, Ma L, Chang D, Coventry A, Gao F, et al. (2014) Neutral genomic regions refine models
19 of recent rapid human population growth. *Proc Natl Acad Sci U S A* 111: 757-762.
- 20 44. Gao F, Keinan A (2016) Explosive genetic evidence for explosive human population growth.
21 *Curr Opin Genet Dev* 41: 130-139.
- 22 45. Wright S (1931) Evolution in Mendelian populations. *Genetics* 16: 97-159.
- 23 46. Ober C, Hyslop T, Hauck WW (1999) Inbreeding Effects on Fertility in Humans: Evidence for
24 Reproductive Compensation. *The American Journal of Human Genetics* 64: 225-231.
- 25 47. Gabriel SE, Brigman KN, Koller BH, Boucher RC, Stutts MJ (1994) Cystic fibrosis heterozygote
26 resistance to cholera toxin in the cystic fibrosis mouse model. *Science* 266: 107-109.
- 27 48. Wagener D, Cavalli-Sforza LL, Barakat R (1978) Ethnic variation of genetic disease: roles of drift
28 for recessive lethal genes. *American Journal of Human Genetics* 30: 262-270.
- 29 49. Quinton PM (1994) Human genetics. What is good about cystic fibrosis? *Curr Biol* 4: 742-743.
- 30 50. Chakravarti A, Chakraborty R (1978) Elevated frequency of Tay-Sachs disease among
31 Ashkenazic Jews unlikely by genetic drift alone. *American Journal of Human Genetics* 30:
32 256-261.
- 33 51. Ewens WJ (1978) Tay-Sachs disease and theoretical population genetics. *American Journal of*
34 *Human Genetics* 30: 328-329.
- 35 52. Richards S, Aziz N, Bale S, Bick D, Das S, et al. (2015) Standards and guidelines for the
36 interpretation of sequence variants: a joint consensus recommendation of the American
37 College of Medical Genetics and Genomics and the Association for Molecular Pathology.
38 *Genet Med* 17: 405-423.
- 39 53. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, et al. (2014) ClinVar: public archive of
40 relationships among sequence variation and human phenotype. *Nucleic acids research* 42:
41 D980-D985.
- 42 54. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE (2015) The Mouse Genome Database (MGD):
43 facilitating mouse as a model for human biology and disease. *Nucleic Acids Res* 43: D726-
44 736.
- 45 55. Haque IS, Lazarin GA, Kang H, Evans EA, Goldberg JD, et al. (2016) MOdeled fetal risk of genetic
46 diseases identified by expanded carrier screening. *JAMA* 316: 734-742.

- 1 56. Mugal CF, Ellegren H (2011) Substitution rate variation at human CpG sites correlates with non-
2 CpG divergence, methylation level and GC content. *Genome Biol* 12: R58.
- 3 57. North BV, Curtis D, Sham PC (2002) A Note on the Calculation of Empirical Values from Monte
4 Carlo Procedures. *The American Journal of Human Genetics* 71: 439-441.
- 5 58. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic
6 features. *Bioinformatics* 26: 841-842.
- 7 59. R Core Team (2015) R: A Language and Environment for Statistical Computing. Vienna, Austria.
- 8 60. Online Mendelian Inheritance in Man O (2016). Baltimore, MD: McKusick-Nathans Institute of
9 Genetic Medicine, Johns Hopkins University.
- 10

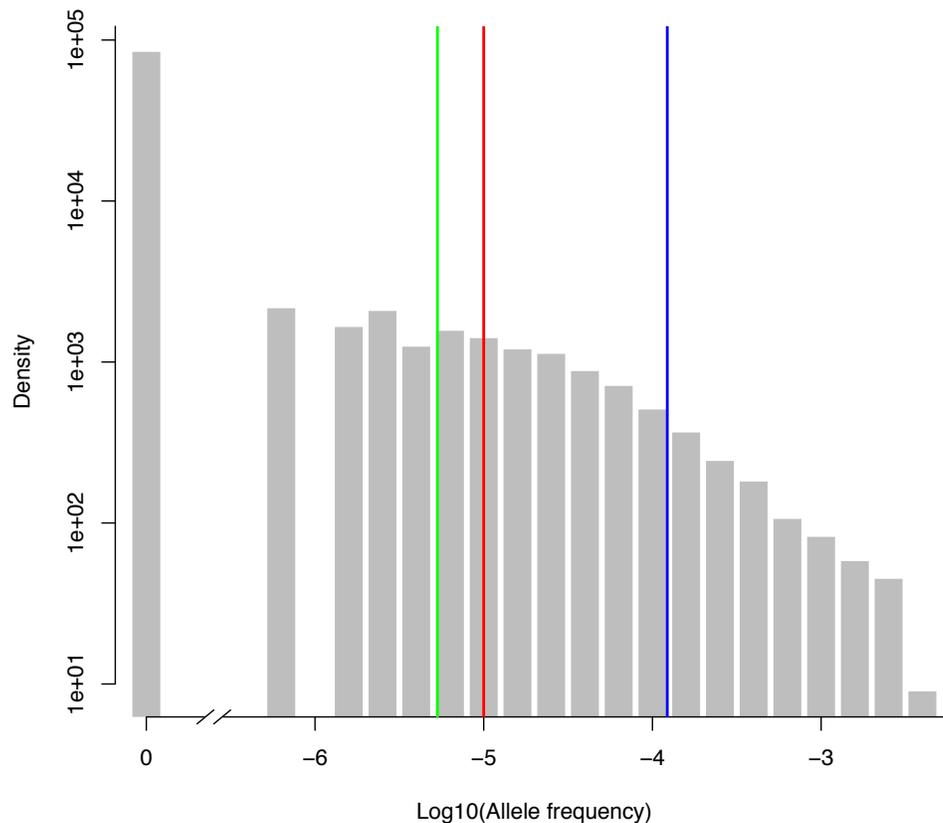
1 **Figures and Tables**

2 **Table 1.** For a given set of parameters n_a (the sample size in a putative disease ascertainment
3 study) and F_a (the average inbreeding coefficient of the population in which the study is being
4 performed), we report the probability of a recessive, lethal mutation being ascertained, as well as
5 the fold increase in mean allele frequencies of the ascertained cases (q_a) in relation to the mean
6 frequencies of all mutations for each mutation type (q_u), based on simulations. See Methods for
7 details. For a similar result derived from analytical modeling, see Fig S8. Parameters for this step of
8 the simulation range according to plausible scenarios for human populations with widespread
9 inbreeding (e.g., $F_a=1/16$ corresponds to offspring of first-cousin marriage). The last column in the
10 bottom panel shows the fold increase of the mean allele frequency observed in ExAC in relation to
11 simulations based on the Tennesen et al. [22] demographic model (see Methods). Mutation rates u
12 per bp, per generation were obtained from a large human pedigree study [16].

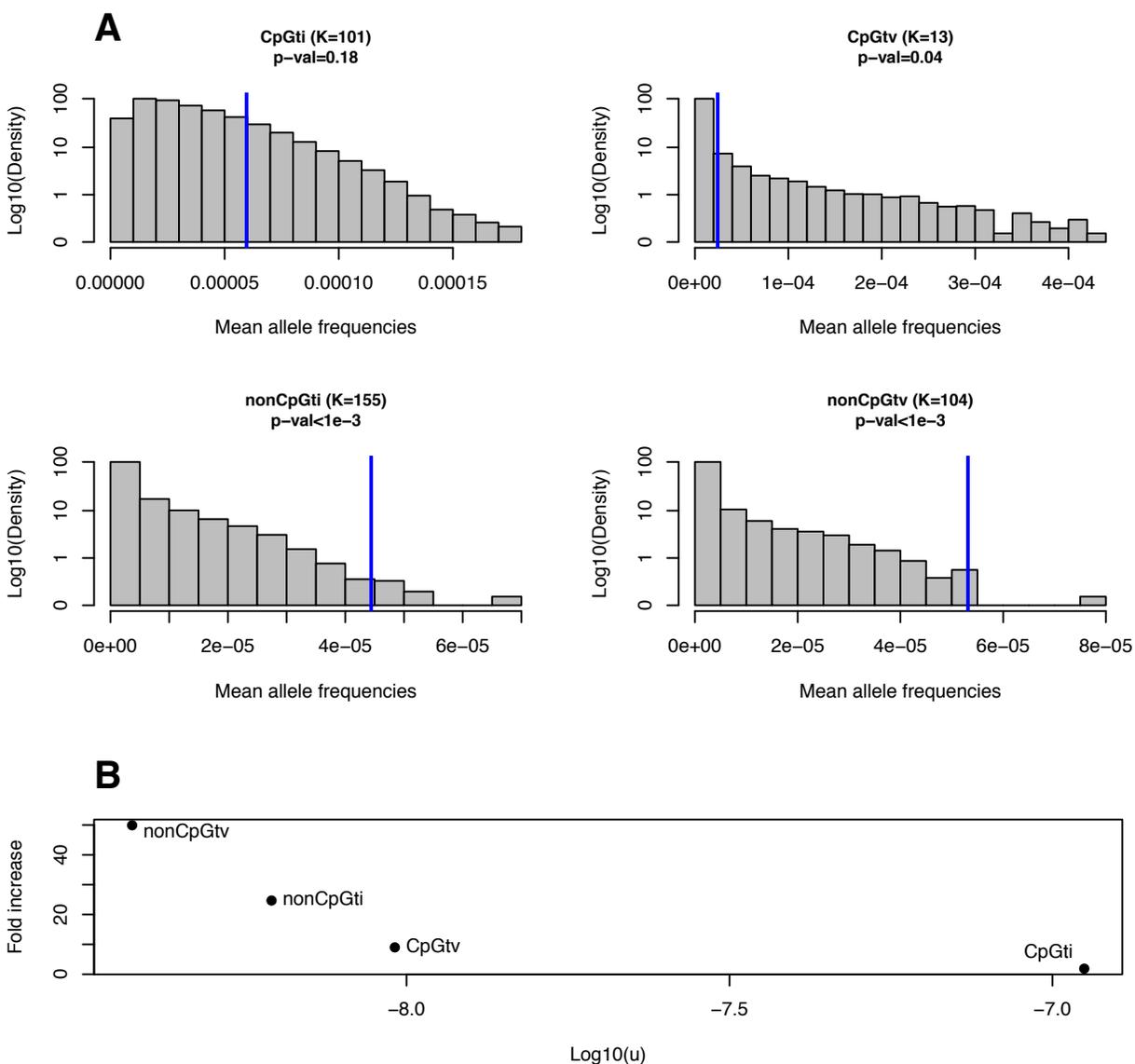
13

Parameter	Simulation settings					
n_a	10,000	10,000	100,000	100,000	500,000	ExAC
F_a	1/16	1/8	1/8	1/6	1/8	
Mutation type (u)	Probability of a mutation being ascertained					
CpGti (1.12e-7)	0.0150	0.0247	0.0968	0.1125	0.1858	-
CpGtv (9.59e-9)	0.0014	0.0024	0.0104	0.0124	0.0235	-
nonCpGti (6.18e-9)	0.0009	0.0015	0.0067	0.0074	0.0146	-
nonCpGtv (3.76e-9)	0.0005	0.0010	0.0044	0.0050	0.0091	-
Fold increase in allele frequency due to ascertainment bias						
CpGti (1.12e-7)	27.6	22.3	9.0	8.0	5.2	1.9
CpGtv (9.59e-9)	284.1	222.6	81.2	70.2	40.6	9.2
nonCpGti (6.18e-9)	455.7	335.6	126.1	114.8	64.7	24.4
nonCpGtv (3.76e-9)	716.5	577.4	192.9	174.1	103.6	49.7

14

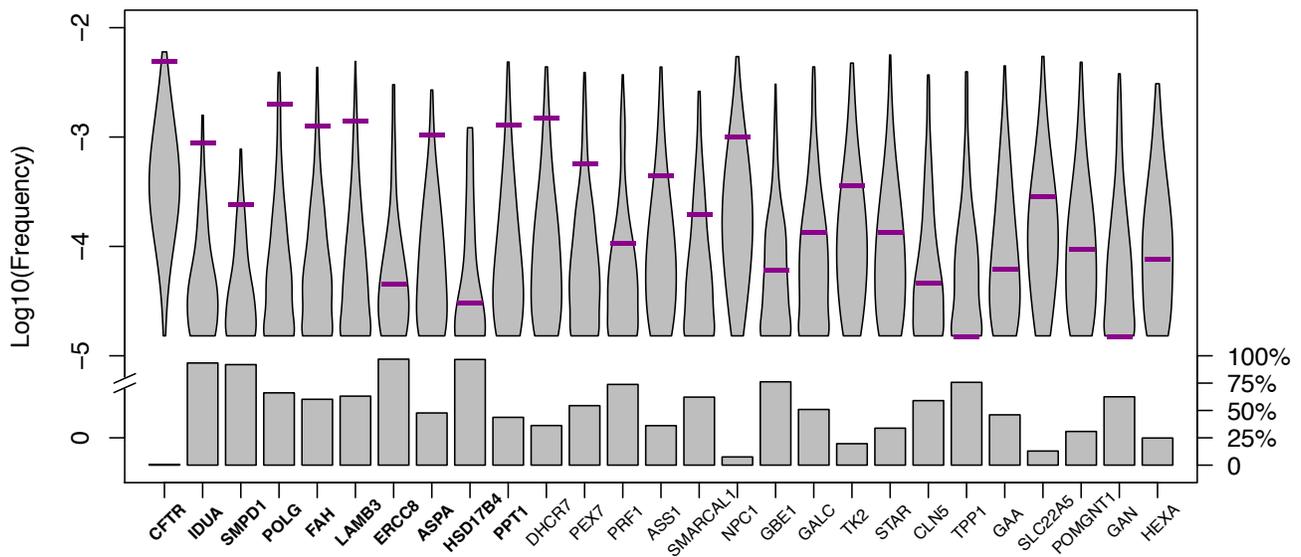


1
2 **Fig 1. Expected allele frequencies in a population, according to mutation-selection balance**
3 **models.** The blue bar denotes the expected allele frequency under an infinite population size, the
4 green bar the mean under a finite constant population, and the red bar the mean under a plausible
5 demographic model for European populations (for this case, the entire distribution across 100,000
6 simulations is shown in the grey histogram). All models assume $s=1$ and $h=0$, i.e. fully recessive,
7 lethal mutations. For the finite constant population size model, we present the mean frequency for
8 a population size of 20,000 (see Fig S2A for other choices). Sample allele frequencies (q) were
9 transformed to $\log_{10}(q)$ and those $q=0$ were set to 10^{-7} for visual purposes, but indicated as “0” on
10 the X-axis. The density of the distribution is plotted on a log-scale on the Y-axis. The mutation rate
11 u was set to 1.5×10^{-8} per bp per generation for all models.



1
 2 **Fig 2. Comparison between expected and observed mean allele frequencies of recessive,**
 3 **lethal disease mutations.** (A) Shown are the expected and observed frequencies of disease
 4 mutations for four different mutation types. The title of the panel indicates the mutation type,
 5 followed by K , the total number of mutations considered in this study of that type, with the p-value
 6 for the difference between observed and expected mean frequencies given below. Distributions in
 7 grey are the mean allele frequencies across K mutations based on 100,000 simulations, and rely on
 8 a plausible demographic model for European populations [22] (see Methods). Blue bars represent
 9 the observed mean frequencies of the four mutation types, estimated from 33,370 individuals of
 10 European ancestry from ExAC. (B) Fold increase in the observed mean allele frequency in relation

- 1 to the expected, as a function of the mutation rate u (on a log-scale), for each of the four mutation
- 2 classes.



1
2 **Fig 3. Disease allele frequencies at the gene level.** The expectation (grey) is based on 1000
3 simulations, assuming no fitness decrease in heterozygotes, but allowing for compound
4 heterozygosity (see Methods for details). The sum of allele frequencies of recessive lethal disease
5 mutations in each gene (purple bars) was obtained from ExAC considering 33,370 European
6 individuals. Genes are ordered according to the two-tailed p-value (Table S4; see Methods). Genes
7 are bolded when they differ significantly from expectation (at the 5% level). Violin plots show the
8 distribution of the \log_{10} allele frequencies among segregating alleles obtained from simulations and
9 boxes represent the fraction of simulations in which no deleterious allele was observed in the
10 simulated sample at present.

Supporting information

Table S1. List of lethal, recessive Mendelian diseases.

Gene	Disease	Phenotype OMIM number	Number of variants that passed "variant filter" ^a
<i>ALS2</i>	Amyotrophic lateral sclerosis 2, juvenile	205100	0
<i>ASPA</i>	Canavan disease	271900	11
<i>ASS1</i>	Citrullinemia, classic	215700	7
<i>BCS1L</i>	Gracile syndrome	603358	1
<i>CFTR</i>	Cystic fibrosis	219700	135
<i>CLN5</i>	Ceroid lipofuscinosis, neuronal, 5	256731	11
<i>DHCR7</i>	Smith-Lemli-Opitz syndrome	270400	17
<i>ERCC8</i>	Cockayne syndrome A	216400	3
<i>FAH</i>	Tyrosinemia, type I	276700	12
<i>FKTN</i>	Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 4	253800	1
<i>GAA</i>	Glycogen storage disease II	232300	21
<i>GALC</i>	Krabbe disease	245200	9
<i>GAN</i>	Giant axonal neuropathy 1, autosomal recessive	256850	4
<i>GBA</i>	Gaucher disease	608013	0
<i>GBE1</i>	Glycogen storage disease IV	232500	6
<i>GNPTAB</i>	Mucopolipidosis II	252500	0
<i>HEXA</i>	Tay-Sachs disease	272800	13
<i>HSD17B4</i>	D-bifunctional protein deficiency	261515	3
<i>IDUA</i>	Hurler-Scheie syndrome Hurler syndrome	607015 607014	12
<i>LAMA3</i>	Epidermolysis bullosa, junctional, Herlitz type	226700	1
<i>LAMB3</i>	Epidermolysis bullosa, junctional, Herlitz type	226700	8
<i>LAMC2</i>	Epidermolysis bullosa, junctional, Herlitz type	226700	0
<i>LARGE</i>	Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 6	613154	0

<i>MCOLN1</i>	Mucopolidosis IV	252650	0
<i>NPC1</i>	Niemann-pick disease, type C1	257220	24
<i>NPHP3</i>	Nephronophthisis 3	604387	0
<i>NPHP4</i>	Nephronophthisis 4	606966	0
<i>OSTM1</i>	Osteopetrosis, autosomal recessive 5	259720	0
<i>PEX7</i>	Rhizomelic chondrodysplasia punctata, type 1	215100	7
<i>PLA2G6</i>	Neurodegeneration with brain iron accumulation 2A	256600	3
<i>POLG</i>	Mitochondrial dna depletion syndrome 4A (Alpers type)	203700	5
<i>POMGNT1</i>	Muscular dystrophy- dystroglycanopathy (congenital with brain and eye anomalies), type A, 3	253280	12
<i>POMT1</i>	Muscular dystrophy- dystroglycanopathy (congenital with brain and eye anomalies), type A, 1	236670	0
<i>POMT2</i>	Muscular dystrophy- dystroglycanopathy (congenital with brain and eye anomalies), type A, 2	613150	0
<i>PPT1</i>	Ceroid lipofuscinosis, neuronal, 1	256730	19
<i>PRF1</i>	Hemophagocytic lymphohistiocytosis, familial, 2	603553	8
<i>SLC22A5</i>	Carnitine deficiency, systemic primary	212140	18
<i>SMARCAL1</i>	Schimke immunoosseous dysplasia	242900	4
<i>SMPD1</i>	Niemann-Pick disease, type A	257200	5
<i>STAR</i>	Lipoid congenital adrenal hyperplasia	201710	9
<i>TCIRG1</i>	Osteopetrosis, autosomal recessive 1	259700	1
<i>TK2</i>	Mitochondrial DNA depletion syndrome 2 (myopathic type)	609560	22
<i>TNFSF11</i>	Osteopetrosis, autosomal recessive 2	259710	0
<i>TPP1</i>	Ceroid lipofuscinosis, neuronal, 2	204500	5
<i>ZMPSTE24</i>	Restrictive dermopathy, lethal	275210	0

^aThis refers to the number of variants in each gene that were associated with the severe form of the corresponding disease, with an early onset reported and no indication of incomplete penetrance or effect in heterozygote carriers (see Methods for details).

Table S2. Information on 417 mutations associated with the severe form of lethal, recessive Mendelian diseases.

Available at <https://github.com/cegamorim/PopGenHumDisease>

Table S3. Information on 91 mutations associated with the severe form of lethal, recessive Mendelian diseases in Counsyl and ExAC databases.

Available at <https://github.com/cegamorim/PopGenHumDisease>

Table S4. P-values for the deviation of expected frequencies of deleterious alleles per gene, in comparison to the expected under SIM.

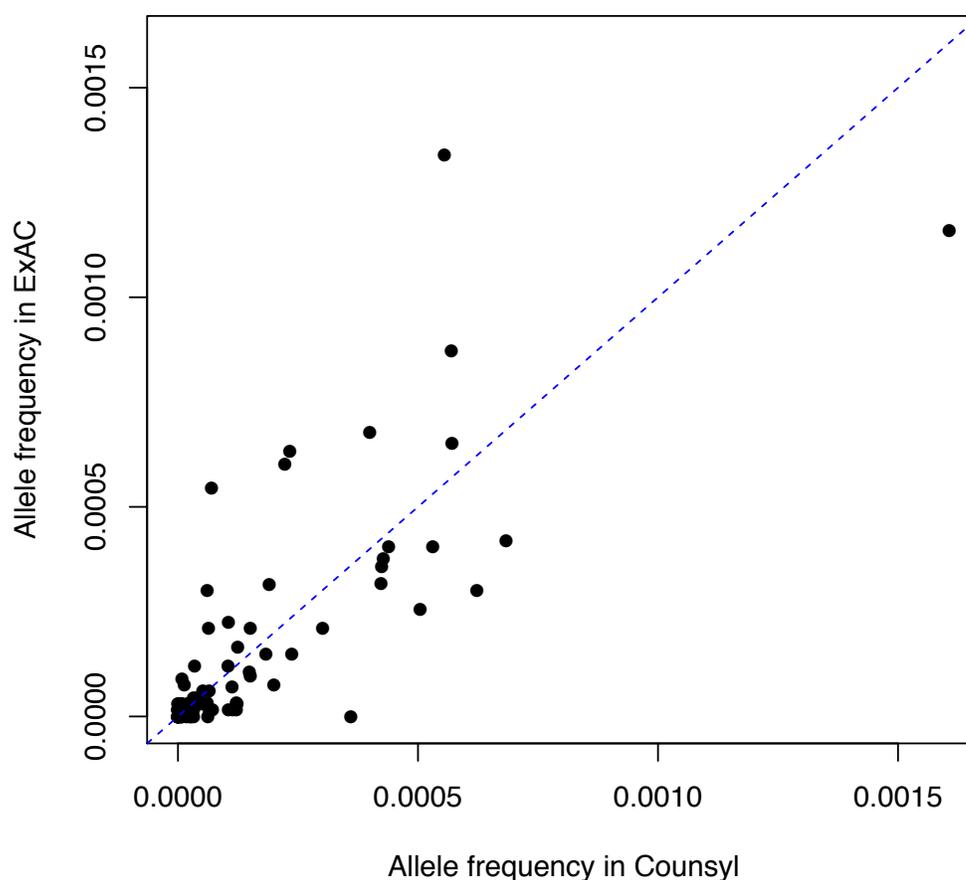
Gene	P-value
<i>ASPA</i>	0.032
<i>ASS1</i>	0.144
<i>CFTR</i>	0.002
<i>CLN5</i>	0.456
<i>DHCR7</i>	0.052
<i>ERCC8</i>	0.03
<i>FAH</i>	0.02
<i>GAA</i>	0.476
<i>GALC</i>	0.3
<i>GAN</i>	0.75
<i>GBE1</i>	0.218
<i>HEXA</i>	0.85
<i>HSD17B4</i>	0.044
<i>IDUA</i>	0.002
<i>LAMB3</i>	0.02
<i>NPC1</i>	0.194
<i>PEX7</i>	0.054
<i>POLG</i>	0.018
<i>POMGNT1</i>	0.62
<i>PPT1</i>	0.048
<i>PRF1</i>	0.132
<i>SLC22A5</i>	0.496
<i>SMARCAL1</i>	0.144
<i>SMPD1</i>	0.01
<i>STAR</i>	0.454
<i>TK2</i>	0.318
<i>TPP1</i>	0.464

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
<i>ASS1</i>	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.
<i>CFTR</i>	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.
<i>DHCR7</i>	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.
<i>NPC1</i>	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, impaired coordination.	Increased brain cholesterol level.

<i>POLG</i>	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.
<i>PRF1</i>	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.	Insulinitis, periinsulinitis, impaired natural killer cell mediated cytotoxicity.
<i>SLC22A5</i>	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.
<i>SMPD1</i>	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 infection, decreased brain size.	Abnormal immune system cell morphology, abnormal neuron differentiation, abnormal depression-related behavior.

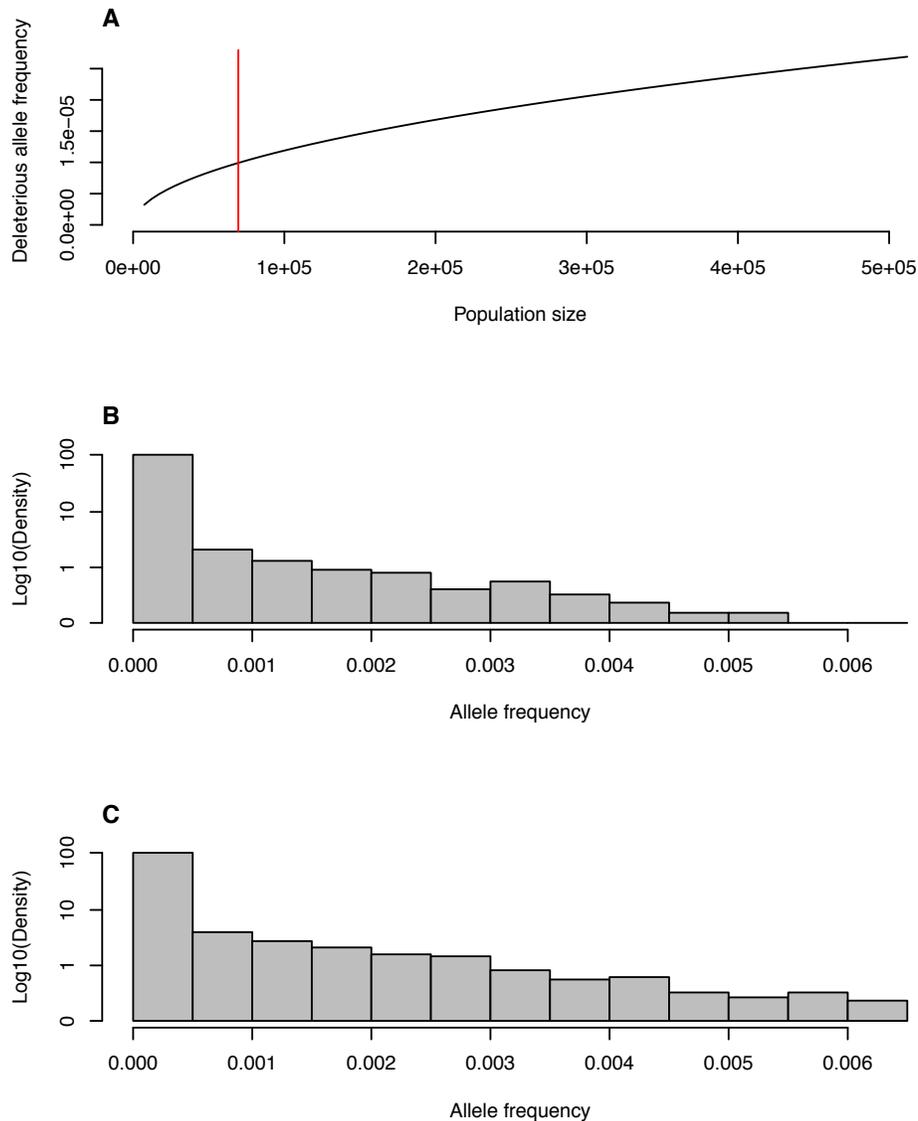
Phenotypes obtained from [60]^a and [54]^b



1

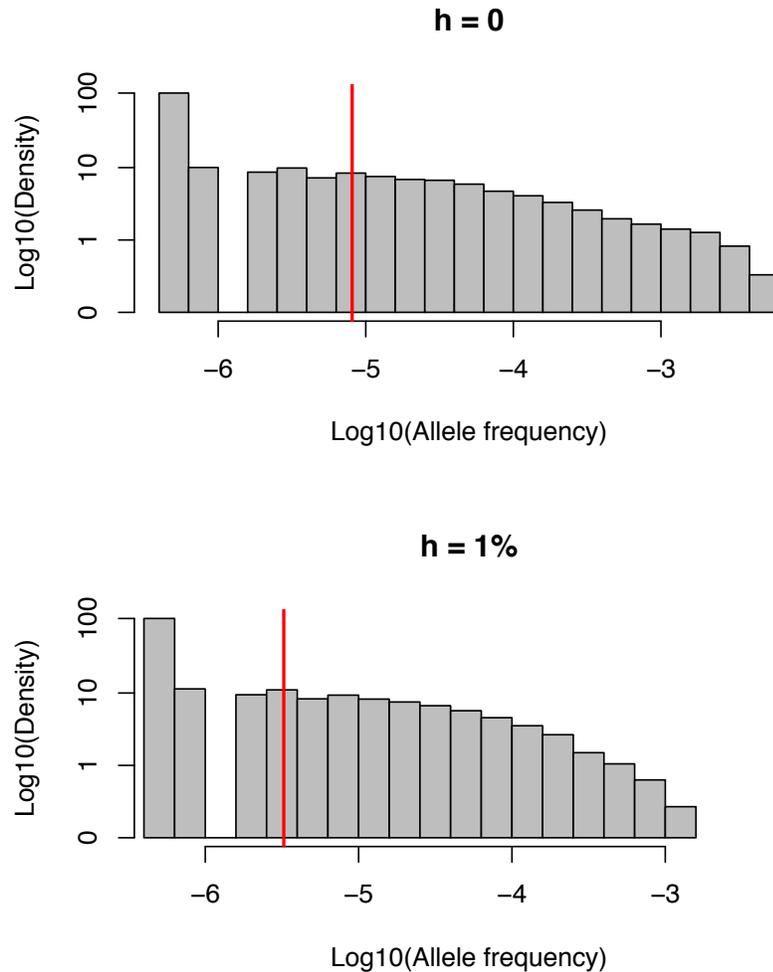
2 **Fig S1. Comparison of the empirical allele frequency of recessive, lethal**
3 **disease mutations in individuals of European ancestry from two large exome**
4 **studies.** Shown are the allele frequencies for 91 variants associated with lethal,
5 recessive diseases, as estimated from 33,370 individuals of non-Finnish, European
6 ancestry in the Exome Aggregation Consortium (ExAC) database [20] and 76,314
7 European-ancestry individuals from a genetic testing laboratory (Counsyl) (see
8 Methods). Dashed blue line indicates cases where allele frequency in Counsyl is the
9 same as in ExAC.

10



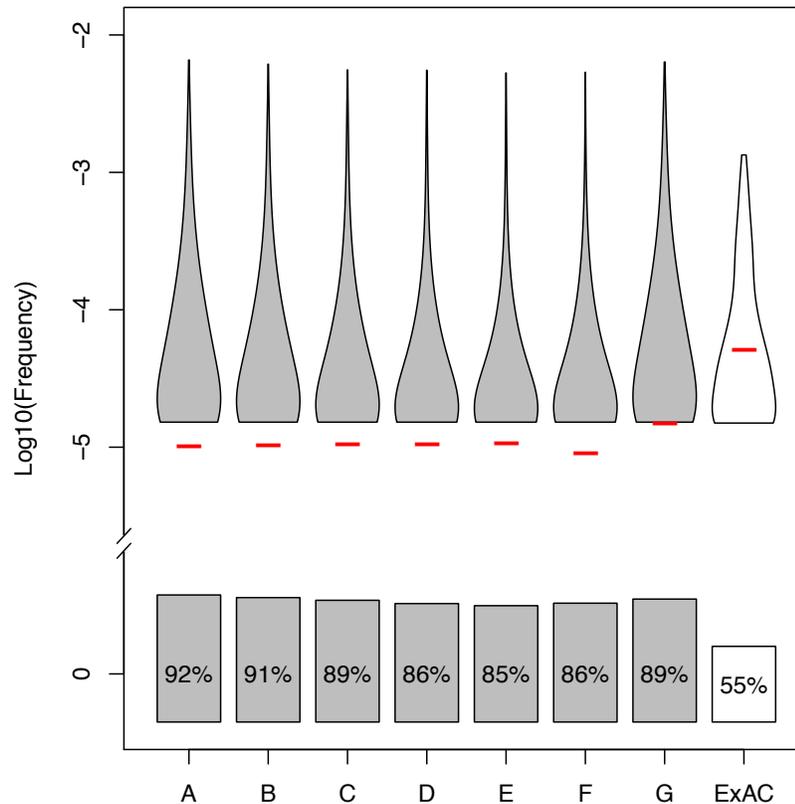
1
2 **Fig S2. Comparisons of mutation-selection balance models for finite**
3 **population sizes.** (A) Mean allele frequency as a function of effective population
4 size, under a model of constant population size. The X-axis range corresponds to the
5 minimum and maximum effective population size estimated in [22]. The red bar
6 indicates the value of a constant population size at which the mean allele frequency
7 predicted is the same as the mean allele frequency estimated in simulations, for an
8 average mutation rate of 1.5×10^{-8} per bp per generation [31]. (B-C) The allele
9 frequency distribution (in grey) is presented for 100,000 observations based on (B)
10 the complex demographic scenario inferred by Tennesen et al. [22] for the

1 evolution of European populations based on simulations (see Methods) and on (C)
2 the finite, constant size population model, with N set to 69,537 individuals to match
3 the mean allele frequency with (B). Both models assume complete lethality ($s=1$)
4 and recessivity ($h=0$).
5



1
2
3
4
5
6
7
8
9
10
11
12

Fig S3. The impact on disease allele frequencies of a small fitness effect in heterozygotes ($h=0.01$). Shown is the distribution of deleterious allele frequencies generated from 100,000 simulations in each case. Means are represented by red vertical bars. For visualization, an allele frequency $q=0$ is set to 0.5×10^{-6} . When a small fitness effect in heterozygotes is considered in the simulations, the mean allele frequency decreases by 61% relative to no effect. The two distributions differ significantly by a Kolmogorov-Smirnov test (p -value $< 10^{-15}$). The mutation rate u was set to 1.5×10^{-8} per bp per generation, reflective of the mean mutation rate estimated for exons [31].



1

2 **Fig S4. Effect of varying the end population size and the average mutation rate**
3 **on the sample frequency of recessive, lethal mutations.** Tennesen et al. [22]

4 inferred the present effective population size of Europeans to be 512,000
5 individuals (column A). We considered the effect of larger present population sizes

6 (2-, 4-, 10-, and 20-fold increase, denoted by columns B, C, D and E respectively),

7 keeping other demographic parameters the same. We also included a model (F)

8 where rapid growth begins immediately after the out-of-Africa bottleneck,

9 representing a more extreme scenario of population growth in comparison to the

10 two-stage and more gradual scenario proposed by Tennesen et al. (2012). For A-F

11 we use the average $u = 1.5 \times 10^{-8}$ per bp per generation [31]. Model G considers a

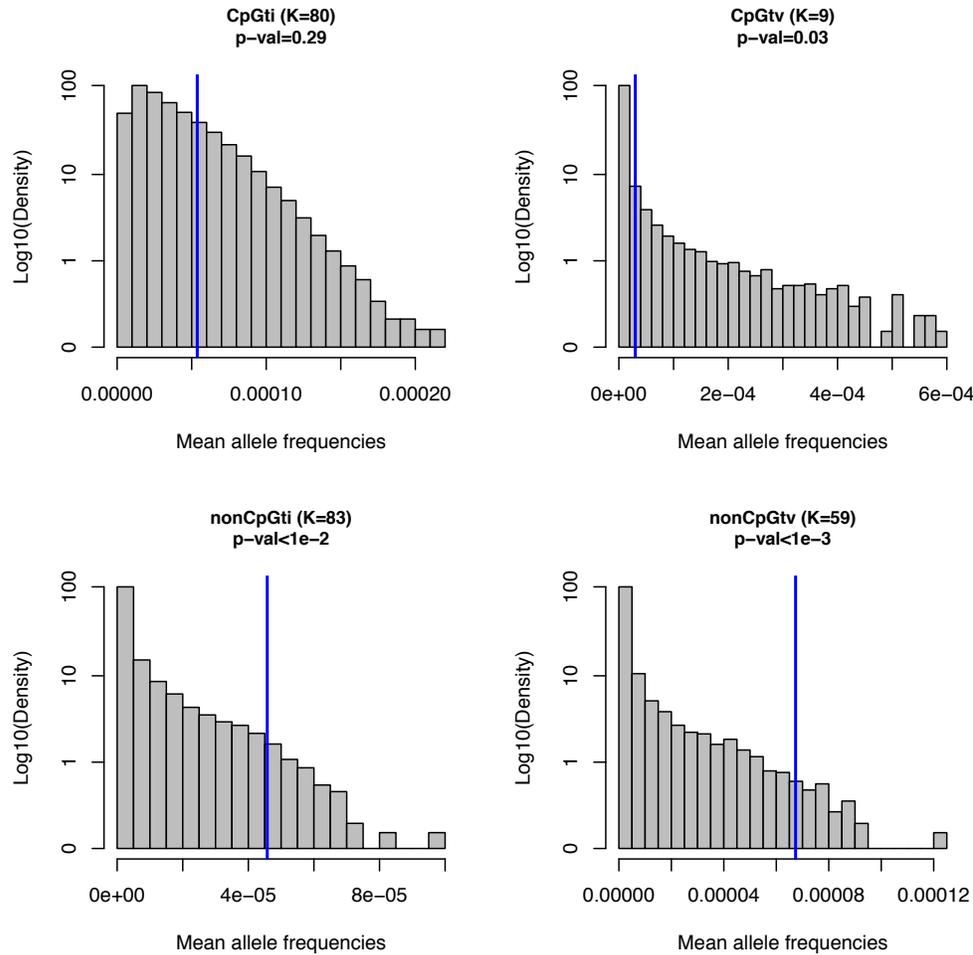
12 larger u (2.25×10^{-8} , i.e., a 1.5-fold increase from A-F), with all other parameters (e.g.,

13 variance in mutation rates across simulations, the demographic model) kept the

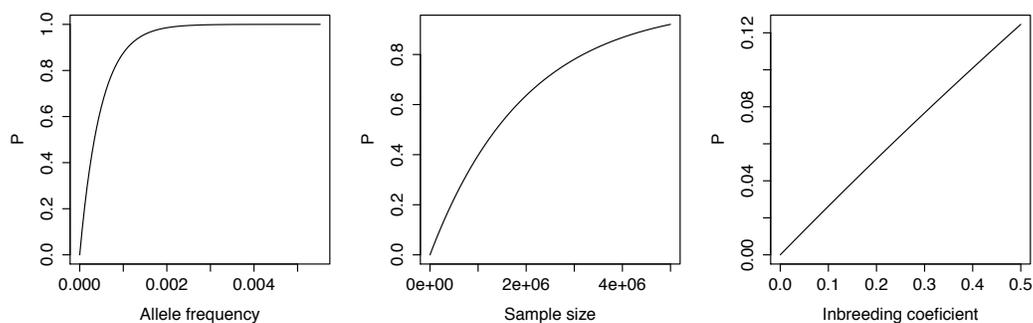
1 same as in column A. The observed allele frequency distribution of 385 disease
2 mutations in ExAC is shown in white. Violin plots show the density distribution of
3 the \log_{10} of the frequency of alleles segregating in these samples, whereas boxes
4 indicate the proportion of sites for which the deleterious mutation was not
5 observed. All distributions differ significantly from one another (i.e., all p-values are
6 $< 10^{-15}$ by a Kolmogorov-Smirnov test).

7

8



1
2 **Fig S5. Expected distribution and the observed mean allele frequencies of**
3 **recessive, lethal disease mutations (excluding mutations in *CFTR* and *DHCR7*).**
4 As in Fig 2, the four panels correspond to four different mutation types. The title of
5 the panel indicates the mutation type, followed by N , the total number of mutations
6 of that type, with p-values for the difference between observed and expected mean
7 frequencies below. Distributions in grey are for 100,000 observations of the
8 expected mean allele frequencies across K mutations, and were obtained from
9 simulations based on a plausible demographic model for European populations [22]
10 (see Methods). Blue bars represent the observed values estimated from 33,370
11 individuals of European ancestry from ExAC. As opposed to in Fig 2, here, we did not
12 include mutations present in two genes (*CFTR* and *DHCR7*) that were outliers in the
13 gene-level analysis (Fig 3) and were reported elsewhere to be carried by healthy
14 homozygous individuals [21].



1

2

3

4

5

6

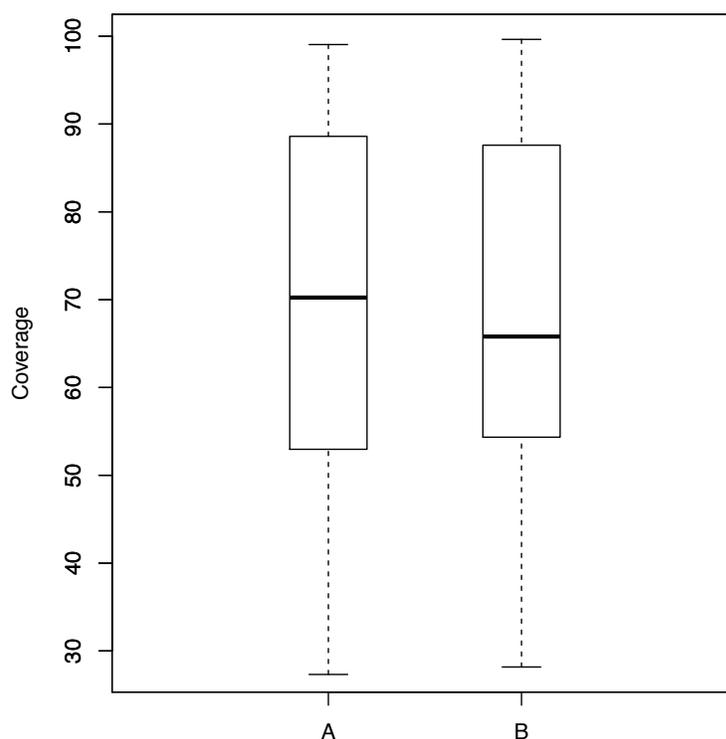
7

8

9

10

Fig S6. The probability of a mutation being ascertained, given its allele frequency p , the sample size n_a of the putative ascertainment study and the inbreeding coefficient F_a in the population in which the ascertainment study was conducted. In each case, we let only one parameter (p , n_a or F_a) to vary, while fixing the others at $p=1 \times 10^{-5}$ (corresponding to the mean allele frequency from simulations), $n_a=10,000$, and $F_a=1/16$ (corresponding to marriage between first cousins, a plausible scenario for a population with widespread inbreeding).



1

2 **Fig S7. Depth of coverage for 385 mutations in ExAC known to cause lethal,**

3 **Mendelian diseases.** Box plots show the mean (black bar) and the lower and upper

4 quartiles for (A) the 248 sites with non-zero sample frequencies in ExAC, for which

5 the number of sequenced non-Finnish European individuals was reported ($n =$

6 32,881) and (B) the 137 sites for which we did not have this information. Since

7 distributions of depth of coverage are similar between the two sets, we assumed

8 that 32,881 individuals were sequenced at all sites, and used this number to

9 subsample simulations to match the sample size of the ExAC data.

10