

# Allele frequencies at recessive disease genes are mainly determined by pleiotropic effects in heterozygotes

Jonathan Judd <sup>1</sup>, Jeffrey P. Spence,<sup>1,2,3,\*</sup> Nikhil Milind <sup>1</sup>, Linda Kachuri,<sup>4,5</sup> John S. Witte,<sup>1,4,5</sup> Jonathan K. Pritchard<sup>1,6,\*</sup>

<sup>1</sup>Department of Genetics, Stanford University, Stanford, CA 94305, United States

<sup>2</sup>Institute for Human Genetics, University of California San Francisco, San Francisco, CA 94158, United States

<sup>3</sup>Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA 94158, United States

<sup>4</sup>Department of Epidemiology and Population Health, Stanford University, Stanford, CA 94305, United States

<sup>5</sup>Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA 94305, United States

<sup>6</sup>Department of Biology, Stanford University, Stanford, CA 94305, United States

\*Corresponding authors: Jeffrey P. Spence, Institute of Human Genetics, University of California San Francisco, San Francisco, CA 94158, United States. Email: jeff.spence@ucsf.edu; Jonathan K. Pritchard, Department of Genetics, Stanford University, Stanford, CA 94305, United States. Email: pritch@stanford.edu

The classic theory of mutation–selection balance predicts the equilibrium frequency of genetic variation under negative selection. The model predicts a simple relationship between the total frequency of deleterious variants, mutation rate, and strength of selection, with different functions for recessive and (co)dominant genes. In this study, we investigate whether genes associated with human recessive disorders fit the predictions of this classic model. By comparing observed frequencies of loss-of-function (LoF) variants to those expected under mutation–selection balance we find that, for nearly all recessive genes, the observed frequencies are too low to be explained by purely recessive selection. Analyzing the effects of heterozygous LoFs on quantitative traits from the UK Biobank, we find that recessive disease genes have widespread quantitative effects in heterozygotes. Together, these results suggest that for a large proportion of recessive disease genes, the majority of selection acting on pathogenic mutations appears to operate through heterozygotes. We conclude that very few human genes follow the classic model of recessive mutation–selection balance.

**Keywords:** population genetics; recessive diseases; quantitative genetics

## Introduction

A century ago, J.B.S. Haldane developed the now-classic theory of “mutation–selection balance,” which predicts the equilibrium frequency of deleterious variants (Haldane 1927, 1937). The key insight of mutation–selection balance is that, at equilibrium, the rate at which new deleterious variants are generated in a gene by mutations must exactly offset the rate at which variants are removed by natural selection.

In the last decade, advances in whole-exome and whole-genome sequencing have enabled sequencing of hundreds of thousands of healthy people (Lek et al. 2016; Karczewski et al. 2020; Chen et al. 2024; Guo et al. 2024). Consistent with the predictions of mutation–selection balance, these studies show that important genes are highly depleted of functional variants such as loss-of-function (LoF) mutations (Lek et al. 2016; Karczewski et al. 2020; Guo et al. 2024). This observation has motivated a great deal of work on estimating the strength of selection against LoF heterozygotes (here denoted  $s_{het}$ ) for each gene, usually assuming additive models of selection (Cassa et al. 2017; Fuller et al. 2019; Agarwal et al. 2023; Zeng et al. 2024).

However, the standard additive model does not apply to fully recessive genes. In this case, selection acts only in homozygotes; since homozygotes are rare compared to heterozygotes, the classic mutation–selection balance theory predicts that pathogenic

variants should be found at much higher frequencies in recessive genes than in dominant or codominant genes.

It is generally believed that the recessive version of mutation–selection balance should apply to a wide variety of autosomal recessive diseases. Recessive diseases, including cystic fibrosis, sickle cell anemia, and Tay–Sachs disease (Loneragan et al. 2001; Fernandes Filho and Shapiro 2004; Ong and Ramsey 2023; McKusick 2024), occur when patients inherit pathogenic mutations that abrogate function in both copies of a particular gene.

While significant progress has been made on estimating  $s_{het}$ , estimating selection coefficients in homozygotes (denoted  $s_{hom}$ ) remains challenging due to the difficulty of disentangling heterozygous effects from homozygous effects in population genetic studies, and due to the rarity of homozygotes and survivorship bias in observational studies (Amorim et al. 2017; Cassa et al. 2017; Balick et al. 2022).

Here, we take a different approach, simply asking whether the frequencies of putative pathogenic mutations in annotated recessive genes are compatible with recessive mutation–selection balance. After addressing several technical challenges we conclude that, for the vast majority of annotated recessive genes, the frequencies of pathogenic mutations are too low to be explained by a fully recessive model.

To explain this, we propose that most recessive disease genes also have meaningful phenotypic effects in heterozygotes—and

that these pleiotropic effects are the main target of selection. At a mechanistic level, we can expect that heterozygous mutations would often reduce gene expression or functionality by 50%; this in turn might affect molecular phenotypes, such as chromatin accessibility, with cascading quantitative effects on organism-level traits (Boyle et al. 2017; Zschocke et al. 2023; Milind et al. 2024; Domingo et al. 2025). Consistent with this prediction, we find that LoF mutations in genes annotated as being fully recessive frequently have significant effects on quantitative traits measured by UK Biobank (UKB) (Barton et al. 2022; Heyne et al. 2023; Fridman et al. 2025). These heterozygous effects are likely too small to be recognized as clinical conditions, and yet strong enough to be subject to negative selection.

In summary, we report here that the frequencies of pathogenic mutations in nearly all annotated recessive genes are too low to be explained by Haldane's classical model of recessive mutation–selection balance. Instead, we propose that most selection at these genes is due to selection acting against quantitative effects in heterozygotes.

## Methods

### Identifying recessive genes

An initial list of 1,183 recessive genes was curated by previous studies (Blekhan et al. 2008; Berg et al. 2013; Amorim et al. 2017) by inclusion in Online Mendelian Inheritance in Man (OMIM), a comprehensive compendium of human genes and genetic phenotypes (Amberger et al. 2015), and manual curation through other resources such PubMed and Gene Reviews (Adam et al. 1993; White 2020). To increase the likelihood that the list of recessive genes only affected traits through autosomal recessive inheritance, the initial list of recessive genes was cross-referenced against the OMIM database in December 2024, and any genes with nonautosomal recessive inheritance were excluded in order to produce the most updated list of 796 genes associated with recessive traits.

### Gene data

Gene-level information about LoF variants, including allele frequencies and mutation locations, were retrieved from the gnomAD database (Karczewski et al. 2020). For this study, we used gnomAD v4.0 genomic information from 705,148 unrelated individuals in the Non-Finnish European (NFE), Finnish (FIN), African (AFR), and South Asian (SAS) ancestry groups. We defined LoFs as those predicted to cause stop-gain, splice-acceptor, and splice-donor effects using the Variant Effect Predictor (VEP; v85) (McLaren et al. 2016). When calculating gene-level LoF allele frequencies ( $p_{\text{LoF}}$ ), we summed allele frequencies for LoFs within the gene, excluding variants with an allele frequency  $>0.005$  to exclude LoFs with potential bioinformatic errors (Zeng et al. 2024; Stolyarova et al. 2025). (Our overall conclusions are robust to excluding this filter.) For each gene,  $p_{\text{LoF}}$  was calculated separately for each population. This variant inclusion/exclusion criteria potentially excludes many pathogenic missense and LoF variants that may lead to a recessive disease in these genes (Molotov et al. 2024). However, the chosen variants are known to be highly likely to stop gene function and increases our confidence in examining pathogenic variation and its effect in recessive mutation–selection balance (Cummings et al. 2020; Karczewski et al. 2020; Stolyarova et al. 2025). Total mutation rates for each gene were calculated by considering all possible LoF-causing point mutations and using a mutation rate that accounts for trinucleotide context and methylation status at CpGs as described in

Cummings et al. (2020) and Karczewski et al. (2020). Final population-specific  $p_{\text{LoF}}$ s and total mutation rates along with the previously determined recessive gene status were used to visualize additive and recessive mutation–selection balance.

### Adjusting mutation–selection balance for missing variation

We wanted to estimate the proportion of pathogenic variants that might be excluded from the types of variants we considered above: stop-gain, splice-donor-disrupting, and splice-acceptor-disrupting point mutations. In particular, we identified the percentage of pathogenic variation in 25 well-known recessive disease genes represented by these classes of variants (Jackson et al. 2018; Xiao and Lauschke 2021). Using ClinVar (Landrum et al. 2025) to identify pathogenic and likely pathogenic short variation in recessive disease genes, we analyzed genes including HBB, CFTR, PAH, ATP7B, MUTYH, HEXA, GBA1, USH2A, and SMN1, which are the genes associated with sickle cell anemia, cystic fibrosis, phenylketonuria (PKU), Wilson disease, MUTYH-associated gastric cancer, Tay–Sachs disease, Gaucher disease, Usher syndrome, and spinal muscular atrophy, respectively (Lonergan et al. 2001; Fernandes Filho and Shapiro 2004; Castellani 2013; Nagral 2014; Garbade et al. 2019; Kumar et al. 2020; Castiglione and Möller 2022; Mercuri et al. 2022; Hemker et al. 2025). After using the ClinVar-provided molecular consequences to identify LoFs and non-LoFs, we calculated the percentage of unique pathogenic variation that were predicted to cause stop gain, disrupt a splice donor, or disrupt a splice acceptor. For each gene, the inverse of these percentages was calculated as a conversion factor to adjust for potential missing non-LoF pathogenic variation, and we adjusted the visualizations of recessive mutation–selection balance by increasing  $p_{\text{LoF}}$  by the average of the calculated conversion factors and the square root total mutation rates by the square root of the average of the conversion factors.

To determine the conversion factors required so that 50% and 90% of recessive disease genes fit the recessive mutation–selection balance model, we identified the recessive disease gene whose ratio between the LoF allele frequency,  $p_{\text{LoF}}$ , and square root total mutation rate,  $\sqrt{\mu L}$ , is at the 50th and 10th percentile of all recessive disease gene ratios, respectively. From the ratios of the genes at the 50th and 10th percentile, the conversion factors are the square of the inverse of the ratio for the gene (Equation (1)).

$$\text{Conversion-Factor} = \left( \frac{1}{\text{ratio}} \right)^2 = \left( \frac{1}{\frac{p_{\text{LoF}}}{\sqrt{\mu L}}} \right)^2 = \left( \frac{\sqrt{\mu L}}{p_{\text{LoF}}} \right)^2 \quad (1)$$

### Mutation–selection balance with inbreeding

Along with the classic mutation–selection balance model (Equation (8)), we extended a model by Nei that calculates the expected equilibrium  $p_{\text{LoF}}$  in a finite population (Supplementary Appendix A) (Nei 1968) to account for the effect of inbreeding depression on recessive genes in a more realistic population (Equation (2)):

$$\mathbb{E}[p_{\text{path}}] = \frac{\sum_{k=0}^{\infty} \frac{(-4N s_{\text{hom}} \phi / (1 + \phi))^k \Gamma(2N\mu / (1 + \phi) + (k + 1)/2)}{k! (2N s_{\text{hom}} (1 - \phi) / (1 + \phi))^{2N\mu / (1 + \phi) + (k + 1)/2}}}{\sum_{\ell=0}^{\infty} \frac{(-4N s_{\text{hom}} \phi / (1 + \phi))^{\ell} \Gamma(2N\mu / (1 + \phi) + \ell/2)}{\ell! (2N s_{\text{hom}} (1 - \phi) / (1 + \phi))^{2N\mu / (1 + \phi) + \ell/2}}} \quad (2)$$

With these two models, we can predict  $p_{\text{LoF}}$  under recessive lethal selection for different total mutation rates and inbreeding values, providing a lower bound for values of  $p_{\text{LoF}}$  consistent with a purely recessive model.

### Simulating populations with varied inbreeding coefficients

In order to assess the fit of Equations (1) and (3) for varying levels of inbreeding assuming recessive lethality, we used SLiM v4.3 to generate simulated population data (Haller and Messer 2023). All simulated populations contained 20,000 individuals and possessed a single recessive lethal gene that accumulated LoF variants over 200,000 generations. We generated 125 simulated populations each with 25 varying and evenly spaced total mutation rates ([0–0.005]) and inbreeding coefficients ([0, 1/64, 1/32, 1/16, 1/8]). After simulating and calculating  $p_{\text{LoF}}$  for each population, we generated theoretical model values for the same parameters using Equations (8) and (2).

The simulated data were plotted over the predicted lower bounds for LoF allele frequency under different levels of inbreeding (Equations (8) and (2)) to assess fit. Additionally, empirical recessive gene data was compared to both models under varying inbreeding conditions to determine the required level of inbreeding for recessive genes to be consistent with purely recessive mutation–selection balance.

### Burden tests

We used REGENIE (Mbatchou et al. 2021) to perform gene-level burden tests on 161 quantitative traits available in UKB (Bycroft et al. 2018). For the whole-genome regression, which is the first step in REGENIE, SNPs from the genotyping array were pruned with a 1,000 variant sliding window with 100 variant shifts and an  $R^2$  threshold of 0.9. SNPs were then filtered to have MAF <1% (to reduce the impact of misannotation and LD with other functional SNPs), genotype missingness <10%, and Hardy–Weinberg equilibrium test  $p$ -value >10<sup>−15</sup>. For LoF burden tests, we used the following covariates: age, sex, age by sex, age squared, 15 genotyping PCs, 20 rare variant PCs, and WES batch. After whole-genome regression, we used the second step of REGENIE to perform burden tests using the whole-exome sequencing data. We used the LOFTEE plugin in Ensembl's VEP to annotate high-confidence LoF sites (McLaren et al. 2016; Karczewski et al. 2020). We further filtered any annotated LoF site with MAF <1% using a variable-frequency filter across genes based on their constraint, as described previously (Milind et al. 2024; Zeng et al. 2024). For each gene, we filtered to individuals that carried at most one LoF variant, ensuring that burden tests were performed using only LoF heterozygotes. Phenotypes were rank-inverse-normal transformed in both the first and second step. The same covariates were used in both steps.

To generate a list of genetically uncorrelated traits, we used previously released estimates of genetic correlation based on common variants from the UKB (<http://www.nealelab.is/uk-biobank/>). Any traits with genetic correlation greater than 0.75 were clustered together, and traits with the most signal based on bias-corrected average squared burden effect size were picked from the clusters to generate a list of 21 uncorrelated traits.

### Measuring burden effects across gene constraint

The estimated heterozygous gene constraint was calculated for all genes to determine whether a positive relationship exists between heterozygous selection and heterozygous trait effects. Prior to constraint estimation, genes falling below the 25th percentile of

total mutation rate were excluded across both recessive and non-recessive genes to reduce noise introduced by genes with limited mutational targets. For nonrecessive genes, heterozygous gene constraint was estimated using GeneBayes-predicted  $s_{\text{het}}$  values, which provide well-calibrated estimates regardless of the observed  $p_{\text{LoF}}$ .

For recessive genes,  $s_{\text{het}}$  was estimated by dividing the gene's total mutation rate by  $p_{\text{LoF}}$  as modeled by additive mutation–selection balance (Equation (5)). Recessive genes with  $p_{\text{LoF}} = 0$  following the mutation rate filter were retained rather than excluded, as neutrality expectations predict at least approximately 10 segregating LoF variants for such genes; observed  $p_{\text{LoF}}$  values near zero therefore likely reflect strong purifying selection rather than an absence of informative data, rendering removal of these genes inappropriate.

Next, for each gene, we estimated the mean squared burden effect size  $\mathbb{E}[\hat{\gamma}^2]$  from the Regenie output. Specifically, because the expectation of the squared estimated effect size satisfies

$$\mathbb{E}[\hat{\gamma}^2] \approx \gamma^2 + \text{SE}^2, \quad (3)$$

an unbiased estimate of the true squared effect size was obtained by subtracting the squared standard error of the test from the squared estimated effect size:

$$\hat{\gamma}_{\text{unbiased}}^2 = \hat{\gamma}^2 - \text{SE}^2. \quad (4)$$

Bias-corrected squared effect sizes were then averaged by gene to provide each gene an estimated constraint value and mean squared burden test effect size. All genes were subsequently segregated into deciles based on estimated gene constraint, and average squared burden test effect sizes within each bin were further stratified by recessive gene status. Final decile effect sizes were separated by recessive classification and plotted to visualize the relationship between heterozygous constraint and heterozygous effect size.

### Sensitivity analysis of the burden enrichment–constraint relationship

To assess the robustness of the observed relationship between heterozygous gene constraint and burden effect size to potential misestimation of  $s_{\text{het}}$  in recessive genes, we performed a sensitivity analysis in which recessive gene  $s_{\text{het}}$  values were scaled by multiplicative factors of  $\times 1$ ,  $\times 0.75$ ,  $\times 0.5$ , and  $\times 0.25$  prior to constraint binning. For each scaling scenario, precalculated  $s_{\text{het}}$  estimates for recessive genes were multiplied by the corresponding factor, and recessive and nonrecessive genes were pooled together to assign constraint decile bins via joint ranking of the scaled recessive and unscaled nonrecessive  $s_{\text{het}}$  values. The nonrecessive reference curve was held fixed by computing its decile assignments from the original unscaled joint distribution ( $\times 1$ ). To reduce noise, the analysis was restricted to genes in the top 75% of all genes by total mutation rate. Loess curves of mean bias-corrected squared burden effect size across constraint deciles were estimated separately for the nonrecessive reference and for recessive genes under each scaling factor.

### Estimating heterozygous selection fraction

To estimate a lower bound on heterozygous gene constraint for recessive genes, we simplified the mutation–selection balance model ( $2\mu_L = 2p_{\text{path}}(1 - p_{\text{path}})s_{\text{het}} + 2p_{\text{path}}^2s_{\text{hom}}$ ) to compute  $s_{\text{het}}$  given the gene's  $p_{\text{path}}$  and total mutation rate, assuming recessive lethality

( $s_{\text{hom}} = 1$ ). Details on the estimator are provided in Supplementary Appendix A. This  $s_{\text{het}}$  estimate serves as the constraint measure for recessive genes in the burden enrichment analysis (Fig. 3c).

Using the same  $s_{\text{het}}$  estimate, we then calculated the proportion of total selection acting through heterozygotes versus homozygotes for each recessive gene, again under the assumption of recessive lethality. This yields an analytical lower bound on the fraction of selection operating through heterozygotes ( $f_{\text{het}}$ ; Fig. 3f), as described in the Results.

## Recessive gene QQ plots for quantitative traits

Quantile–quantile plots were generated for all recessive genes using  $-\log_{10}(p\text{-values})$  generated from LoF burden tests for a list of 161 UKB quantitative traits and subset of 21 uncorrelated quantitative traits. Observed  $-\log_{10}(p\text{-values})$  were compared to expected values from a uniform distribution. 95% confidence intervals were generated from a beta distribution using the 2.5% and 97.5% for the  $i$ th smallest expected  $p$ -value of the uniform distribution.

## Aggregated Cauchy association tests

For each recessive disease gene, Aggregated Cauchy Association Test (ACAT) was used to combine all LoF burden test  $p$ -values for the gene (Liu et al. 2019). We performed this analysis on different sets of burden test results: (1) combining LoF burden tests for all 161 UKB quantitative traits, and (2) combining LoF burden tests for 21 uncorrelated quantitative traits. This analysis would identify recessive genes with significant nonrecessive effects using the available trait data. This analysis was repeated for all 18,398 nonrecessive genes independent of our curated recessive disease gene list. With all the nonrecessive genes, ACAT was used to combine all LoF burden test  $p$ -values for each gene across the LoF burden tests for all 161 UKB quantitative traits and LoF burden tests for 21 uncorrelated quantitative traits.

## Genomic characterization of recessive disease genes

To assess whether recessive disease genes are distinguished from nonrecessive protein-coding genes by their genomic location or functional features, we performed two complementary analyses. In the first, we tested whether the 796 recessive disease genes are nonrandomly distributed across autosomes relative to all protein-coding genes in the analysis set. For each autosome, the expected number of recessive genes was computed under the null hypothesis that recessive genes are distributed proportionally to the total number of protein-coding genes per chromosome. We assessed overall deviation from proportionality using a chi-squared goodness-of-fit test and tested per-chromosome enrichment or depletion using two-sided Fisher's exact tests with Bonferroni correction for 22 tests. Chromosome assignments for each gene were obtained from the *org.Hs.eg.db* Bioconductor annotation package using Ensembl gene identifiers.

In the second analysis, we compared recessive and nonrecessive genes across a panel of genomic features spanning gene constraint, gene expression, promoter sequence composition, and chromatin environment (Supplemental Table 4). Gene-level total mutation rates and  $s_{\text{het}}$  estimates were derived from gnomAD v4.0 as described above. Gene expression metrics were obtained from the GTEx v8 release, which reports median gene-level transcripts per million (TPM) across 54 tissues from 948 postmortem donors (The GTEx Consortium 2020). For each gene, we computed the maximum and mean of the median TPM values across all tissues. GC content was calculated for the promoter region (TSS  $\pm 2$  kb) of

each gene from the GRCh37 reference genome sequence. Chromatin environment features were derived from the Roadmap Epigenomics Consortium 15-state ChromHMM model for the E003 reference epigenome (H1 human embryonic stem cells) (Kundaje et al. 2015). Promoter activity was defined as the base-pair-weighted fraction of the TSS  $\pm 2$  kb window annotated as active promoter states (TssA, TssAFlnk). DNase-seq chromatin accessibility was quantified as the mean  $-\log_{10}(p\text{-value})$  signal from the consolidated DNase-seq track within the same promoter window. Enhancer density was defined as the fraction of the gene body  $\pm 10$  kb annotated as enhancer states (EnhG, Enh). Differences in feature distributions between recessive and nonrecessive genes were assessed using Wilcoxon rank-sum tests. Within recessive genes, we tested Spearman rank correlations between each feature and the estimated proportion of total selection acting through heterozygotes. All  $p$ -values were Bonferroni-corrected for the number of features tested within each analysis.

## Identifying and analyzing purely recessive genes

Recessive genes that possessed a  $p_{\text{LoF}}$  greater than their respective square root total mutation rate were identified to follow classic recessive mutation–selection balance and potentially only be affected by homozygous selection. As population-specific  $p_{\text{LoF}}$  were calculated, the list of “purely” recessive genes varied by population and were individually identified.

Gene Ontology (GO) enrichment analysis was performed on recessive genes identified as “purely” recessive in the NFE, FIN, AFR, and SAS populations to identify shared biological, molecular, or cellular function among identified recessive genes (Ashburner et al. 2000; The Gene Ontology Consortium et al. 2023).

## Mutation–selection balance with balancing selection

To understand the potential role of balancing selection on the mutation–selection balance for genes associated with recessive diseases, we visualized the relationship between  $p_{\text{LoF}}$  and square root total mutation rate for genes previously suspected to be affected by balancing selection and are associated with only recessive diseases (Clark 1998; Risch et al. 2003; Andrés et al. 2009; Alfonso-Sánchez et al. 2010; Garbade et al. 2019). We performed this analysis using observed  $p_{\text{LoF}}$  from NFE, FIN, AFR, and SAS populations available from gnomAD v4.0 genomic information.

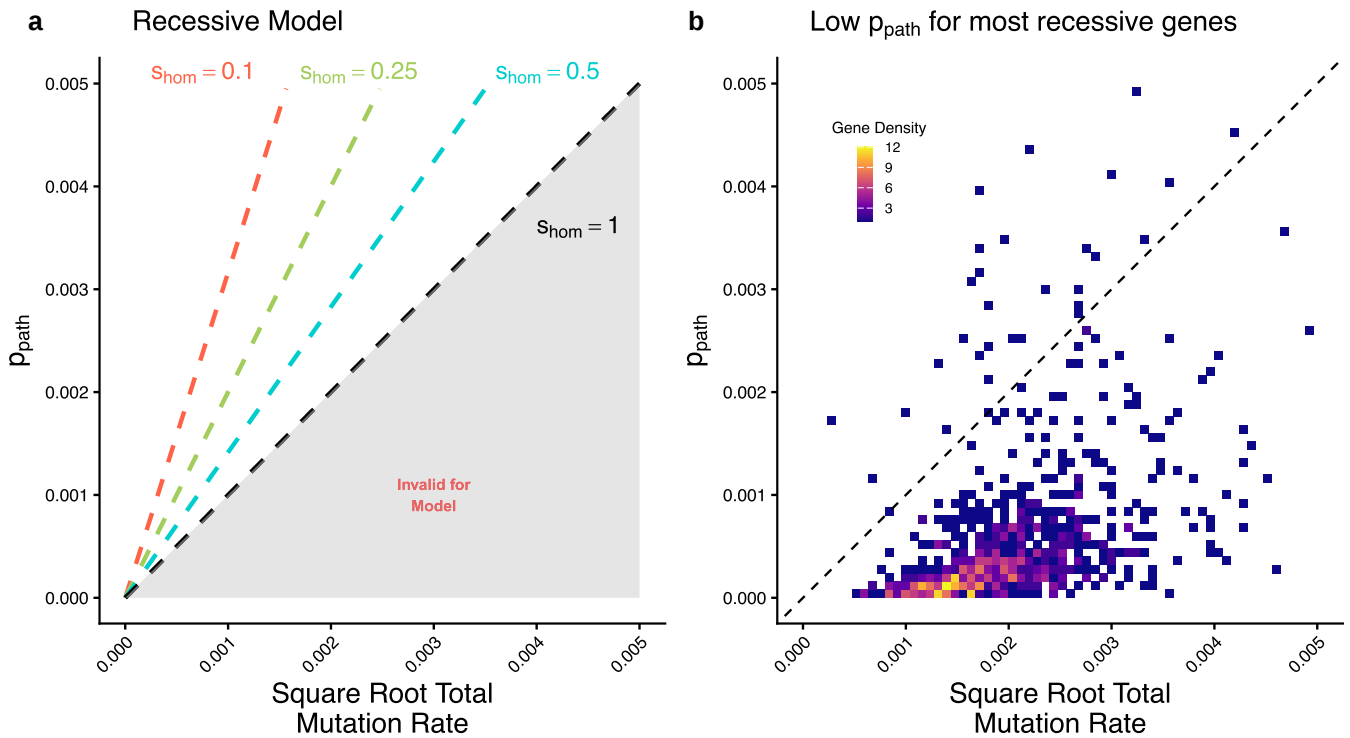
## Results

Mutation–selection balance predicts the total allele frequency of all pathogenic variants for a gene,  $p_{\text{path}}$ , by calculating the frequency at which the rate of new pathogenic alleles arising by mutation is exactly offset by the removal of alleles by natural selection:

$$2\mu L = 2p_{\text{path}}(1 - p_{\text{path}})s_{\text{het}} + 2p_{\text{path}}^2s_{\text{hom}}, \quad (5)$$

where  $\mu$  denotes the per-base pair mutation rate and  $L$  is the effective number of sites that can produce pathogenic mutations (Haldane 1927, 1937). When selection on heterozygotes,  $s_{\text{het}}$ , is non-negligible (i.e. when mediated by dominant or codominant effects on traits), nearly all selection occurs in heterozygotes and the equilibrium frequency can be approximated as

$$p_{\text{path}} = \frac{\mu L}{s_{\text{het}}}; \quad s_{\text{het}} \gg 0. \quad (6)$$



**Fig. 1.** Observed pathogenic allele frequency versus expected: a) theoretical relationship between the square root mutation rate and  $p_{\text{path}}$  under recessive mutation–selection balance, for a range of values of homozygous selection,  $s_{\text{hom}}$ . The shaded area represents parameter combinations that are incompatible with classic recessive mutation–selection balance. b) Density plot showing the observed relationship between square root total mutation rate and predicted  $p_{\text{path}}$  for recessive genes. Most genes lie in the parameter region that is inconsistent with classic mutation–selection balance.

In contrast, for fully recessive genes, selection occurs only in homozygotes, so the equilibrium frequency is determined entirely by the frequency of homozygotes:

$$p_{\text{path}} = \sqrt{\frac{\mu L}{s_{\text{hom}}}}; \quad s_{\text{het}} \approx 0. \quad (7)$$

Under this model, the frequency of pathogenic alleles,  $p_{\text{path}}$ , is proportional to the square root of the gene's total mutation rate  $\mu L$  (Fig. 1a). A key prediction for the present paper is that, even for fully lethal recessive mutations ( $s_{\text{hom}} = 1$ ), selection cannot push the equilibrium frequency,  $p_{\text{path}}$ , below  $\sqrt{\mu L}$ .

Note that this classic prediction makes several simplifying assumptions that may not hold in practice (i.e. ignoring inbreeding, drift, nonequilibrium population dynamics, and in-phase compound heterozygotes). We will consider these below.

**Curation of recessive disease genes.** To assess the extent to which purely recessive mutation–selection balance pertains to human genes, we curated a list of genes associated with only recessive diseases. To identify the most plausible purely recessive genes, we intersected the 19,194 protein-coding genes defined from the HUGO Gene Nomenclature Committee (Seal et al. 2023), with previous lists of recessive disease genes (Blekhrman et al. 2008; Berg et al. 2013; Amorim et al. 2017). We then further filtered to remove genes with nonrecessive effects reported in the OMIM database (Amberger et al. 2015). This removes 387 genes, such as *CFTR* (cystic fibrosis) and *HBB* (sickle cell anemia), that are annotated as recessive, but also have documented heterozygous or nonrecessive effects.

This results in a final list of 796 genes with the strongest support for a fully recessive mode of inheritance

(Supplementary Table 1). To assess whether this gene set is genomically representative of protein-coding genes in general, we examined the chromosomal distribution and genomic features of recessive disease genes relative to all nonrecessive protein-coding genes included in the analysis. Recessive disease genes show no significant enrichment or depletion on any autosome relative to expectations based on protein-coding gene density (Supplementary Table 2), and they are indistinguishable from nonrecessive genes with respect to gene expression, chromatin accessibility, promoter activity, enhancer density, and promoter GC content (Supplementary Table 3). The only features that differ significantly between the two sets are total mutation rate and estimated  $s_{\text{het}}$ . While the  $s_{\text{het}}$  result is expected due to previous observations (Zeng et al. 2024) and the given framework employed here, the difference in total mutation rate is unexpected and worthy of further investigation. These results indicate that the deviation from recessive mutation–selection balance observed in subsequent analyses is unlikely to reflect systematic genomic differences between annotated recessive and nonrecessive genes.

**Comparison of  $p_{\text{path}}$  to model predictions.** Previous studies of selection coefficients in heterozygotes ( $s_{\text{het}}$ ) have generally focused on estimating the frequencies of certain types of LoF mutations that can be accurately predicted computationally, and for which it is relatively straightforward to model the corresponding mutation rates. Specifically, recent studies have focused on annotating stop-gain, splice site donor-/acceptor-disrupting variants (Cummings et al. 2020; Karczewski et al. 2020).

Here, we refer to these specific categories of mutations as LoF mutations, and we denote the aggregate frequency of LoF variants as  $p_{\text{LoF}}$ . Gene-level information about LoF variants, including positions and allele frequencies, were retrieved from the gnomAD database v4.0, calculated separately for the NFE, FIN, AFR, and

SAS, ancestry groups (Cummings et al. 2020; Karczewski et al. 2020; Stolyarova et al. 2025). We excluded variants with an allele frequency of  $>0.005$  as previous work has shown that these are highly enriched for LoFs that are either bioinformatic errors, or may otherwise be functionally rescued (Zeng et al. 2024). (We show below that our qualitative results hold even without this filtering.)

We estimated the gene level mutation rate  $\mu_{L_{LoF}}$  (to LoFs) for each gene by counting all possible LoF-causing point mutations. We then computed a context-specific mutation rate that accounts for trinucleotide context and methylation status at CpGs for each possible LoF mutation; this was averaged across all  $L$  possible LoF-generating mutations, to estimate a total gene-level mutation rate (Cummings et al. 2020; Karczewski et al. 2020).

Initial analysis showed that, for nearly all recessive disease genes, the frequency of LoFs,  $p_{LoF}$ , lies far below the expected minimum frequency of  $\sqrt{\mu_{L_{LoF}}}$  indicating that  $p_{LoF}$  cannot be predicted by mutation–selection balance under a purely recessive model (Supplementary Fig. 1a). However, this calculation fails to capture the fact that many individuals who are impacted by recessive disease are compound heterozygotes, carrying different mutations on each of their two gene copies. Thus, strictly speaking, the theory only applies to the full set of all possible pathogenic mutations,  $p_{path}$ .

Compared to annotating LoF mutations, it is much more difficult to confidently identify all pathogenic mutations for a given gene, let alone to estimate the total mutation rate that could generate pathogenic mutations.

As an alternative, we estimated a scaling factor that represents the fraction of all pathogenic mutations captured by LoFs. Specifically, we identified the molecular consequences of all known pathogenic variants for 25 well-studied recessive disease genes (Supplementary Fig. 1b) and then calculated the percentage of this variation that is captured by our annotated LoF mutations (Supplementary Fig. 1c and Supplementary Table 4). We found that, on average, LoFs contributed around 32% of the total number of pathogenic variants. We then used the inverse of this fraction as a conversion factor to rescale  $p_{LoF}$  and the LoF mutation rate  $\mu_{L_{LoF}}$  to estimate an overall  $p_{path}$  and total mutation rate  $\mu_L$  for each gene.

We were intrigued to find that, for the vast majority of recessive disease genes, the estimated  $p_{path}$  is far lower than can be explained by the recessive mutation–selection balance model (Fig. 1b). This result is observed when examining the relationship between  $p_{path}$  and total mutation rate in recessive disease genes in other populations (Supplementary Fig. 2).

When the list of recessive genes is restricted to recessive-lethal genes, which are under the strongest homozygous selection, we also see quantitatively similar results (Supplementary Fig. 3a). This subset represents the most extreme proxy for disease severity available in current gene annotations, and the poor fit of even these genes to the purely recessive model indicates that the deviations we observe are not confined to conditions with modest fitness consequences.

Additional sensitivity analyses were performed to confirm that recessive disease gene selection and variant selection criteria, which might result in underestimating  $p_{path}$ , were not likely factors to explain these observations. We tested whether the data might fit if we removed the filter on unusually high-frequency LoFs (MAF  $>0.005$ ), but this makes little difference to the qualitative results (Supplementary Fig. 3b). We also wondered whether these observations might be explained if we are underestimating the ratio of LoFs to total pathogenic mutations. However,

we find that we would need to increase this factor by around 104-fold to make 50% of genes fit the model and need to increase this factor an implausible amount (1,465-fold) to make 90% of genes fit the model.

To better understand why recessive disease genes are largely inconsistent with recessive mutation–selection balance, we next considered two potential categories of explanations: deviations from the basic model including inbreeding and drift, and unexpected quantitative effects in heterozygotes.

## Potential model deviations

The classic mutation–selection balance model makes three important simplifications that may not hold in practice: no inbreeding, no drift, and modeling the gene essentially as a single site with high mutation rate. We consider each of these in turn.

**Inbreeding.** In our standard model, the probability that an individual is homozygous for pathogenic mutations follows Hardy-Weinberg expectations, i.e.  $p_{path}^2$ . However, many populations exhibit some degree of inbreeding—i.e. relatives mate with one another more often than expected under purely random mating. Close inbreeding dramatically increases the probability that an offspring inherits a pathogenic mutation from both parents (Fig. 2a). Thus, conditional on allele frequency, inbreeding increases the probability that pathogenic mutations will be exposed to selection; this results in lower equilibrium frequencies compared to random mating (Charlesworth and Willis 2009). We wanted to assess whether this effect is strong enough to explain our results.

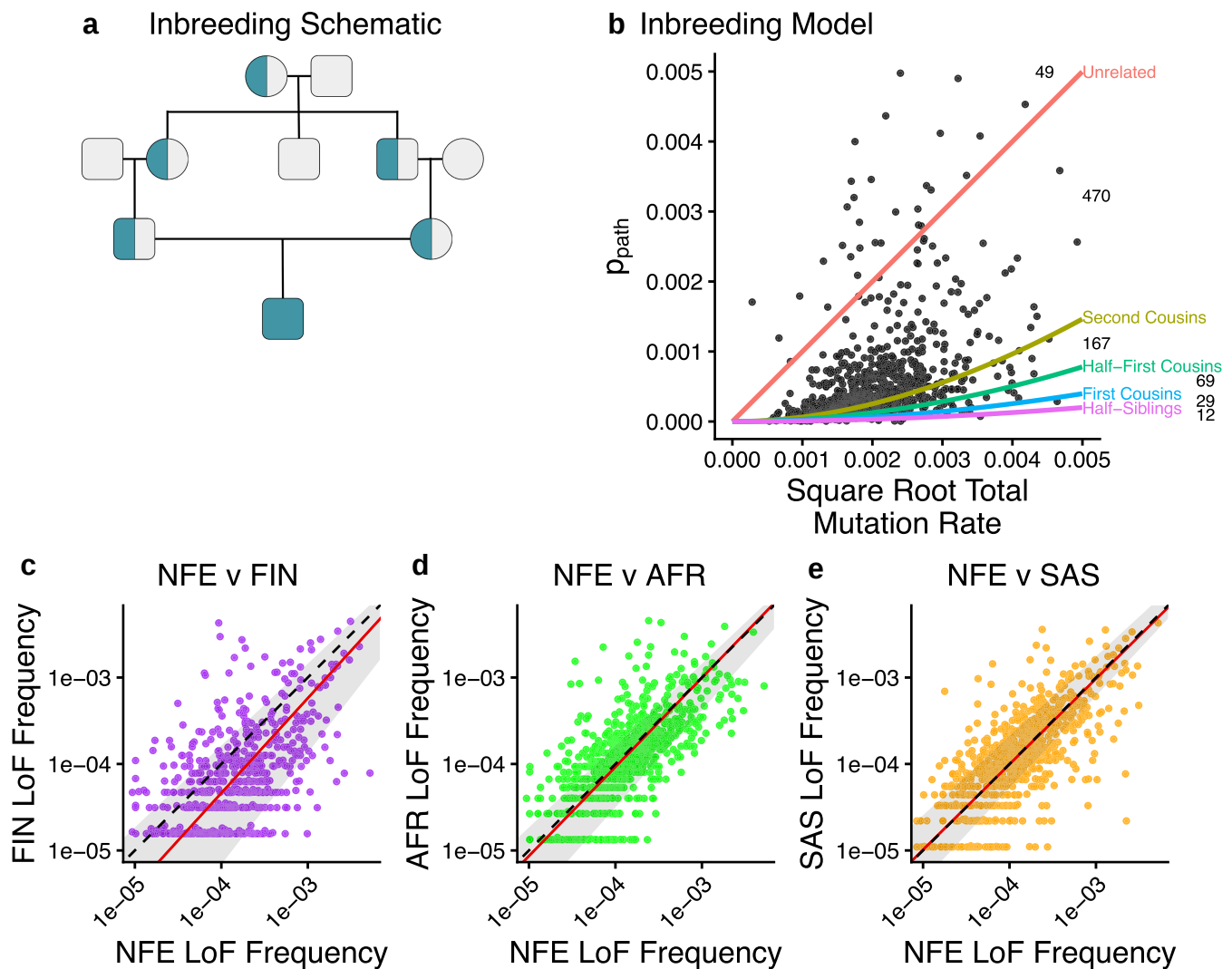
To model this, we measure the extent of inbreeding using the probability that an individual inherits both of their copies of a gene identically by descent, which we denote  $\phi$ . For example, under random mating  $\phi = 0$ ; if 20% of matings in a completely outbred population are between first cousins then  $\phi = 0.0125$ . Under this model, mutation–selection balance takes the following form in the large population size limit:

$$\lim_{N \rightarrow \infty} \mathbb{E}[p_{path}] = -\frac{\phi}{2(1-\phi)} + \sqrt{\left(\frac{\phi}{2(1-\phi)}\right)^2 + \frac{\mu_L}{(1-\phi)s_{hom}}} \quad (8)$$

See Supplementary Appendix A for details of the model and the derivation of the results.

Equation (8) highlights that higher inbreeding coefficients reduce the expected  $p_{path}$  for a given total mutation rate. However, the extent of inbreeding would need to be implausibly high to explain the data: to explain 95% of recessive genes, we would need a level of inbreeding equivalent to a single generation where all matings occur between first cousins (Fig. 2b and Supplementary Fig. 4a).

For comparison, some contemporary inbreeding rates (estimated from 1,000 Genomes data Durbin et al. 2010) identify low levels of second cousin inbreeding in NFE populations and no evidence of first-cousin inbreeding (Gazal et al. 2015) while others (estimated from UK Biobank Data) similarly only identify low rates of second- and first-cousin inbreeding in specific European populations (Yengo et al. 2019). While repeated generations of inbreeding could, in principle, cause identity-by-descent to accumulate beyond the levels implied by a single-generation model, as individuals who already share excess identity-by-descent may preferentially mate within such communities, this effect is unlikely to account for the implausibly high inbreeding coefficients that would be required to fit the observed data. Overall, these results indicate that inbreeding depression is unlikely to



**Fig. 2.** Inbreeding under mutation–selection balance: a) illustration of how inbreeding, in this case between first cousins, increases the frequency of homozygotes in a population (conditional on  $p_{\text{path}}$ ). b) The lines show the theoretical equilibrium  $p_{\text{path}}$  under different levels of inbreeding accounting for the potential effects of genetic drift; these are superimposed onto the observed data (e.g. dot is a single recessive gene). The numbers indicate the number of genes that lie between each adjacent pair of theoretical lines. Comparison of  $p_{\text{LoF}}$  of recessive genes between NFE and c) FIN populations, d) AFR populations, or e) SAS populations, respectively. Red line represents the total least squares regression of  $p_{\text{LoF}}$  between populations. Shaded region represents 95% confidence interval of regression line. The black lines indicate  $y = x$ .

explain the failure of recessive mutation–selection balance to fit recessive disease genes.

*Drift in small populations.* Equation (8) assumes an infinite-sized population, neglecting the effects of genetic drift on  $p_{\text{path}}$ . In contrast to the additive case, genetic drift lowers the expected  $p_{\text{path}}$  under mutation–selection balance (Nei 1968; Laws and Jamieson 2011; Hedrick and Garcia-Dorado 2016; Robinson et al. 2018).

To address this, we derived an approximation that quantifies the effects of drift, inbreeding, and selection in homozygotes (Supplementary Appendix A). Our approximation provides a good fit to forward-in-time Wright-Fisher simulations (Supplementary Fig. 4b). Under this adjusted model, inbreeding has a greater effect in reducing  $p_{\text{path}}$ , which enables more recessive disease genes to fit mutation–selection balance for lower levels of inbreeding. However, even when accounting for drift, fitting the data for most genes would still require inbreeding on the level of ubiquitous first cousin mating (Supplementary Fig. 4c and d). We also argue in Supplementary Appendix A that

violations of the assumption that populations are at equilibrium are unlikely to alter these findings.

We also reasoned that if either of these demographic factors (inbreeding or drift) is somehow much larger than we expect, then it would likely vary in magnitude across populations. We would not expect that the precise parameters are the same in diverse populations. But when we analyze the fit for  $p_{\text{path}}$  in diverse human populations we find only slight variation among populations and no clear bias (Fig. 2c–e), a result previously described by Stolyarova et al. (2025). While there is evidence of a higher value of  $\phi$  in SAS populations (Gazal et al. 2015; Wall et al. 2023), this does not result in the systematic reduction of  $p_{\text{LoF}}$  relative to populations with lower levels  $\phi$  that we would expect to see under a purely recessive model. The quantitative similarities among diverse populations suggest that a biological explanation is more likely than a demographic explanation.

*In-phase compound heterozygotes.* Lastly, the standard model does not precisely model compound heterozygotes.

These occur when an individual inherits a pair of pathogenic mutations at different sites within the same recessive gene. Typically, a compound heterozygote would have the disease if the two mutations are out-of-phase (i.e. damaging both gene copies), but not if the mutations are in-phase (i.e. damaging only one gene copy). Since selection does not act against in-phase compound heterozygotes, the net effect of this is to weaken selection, and increase the expected  $p_{\text{path}}$  (Guo et al. 2024; Lassen et al. 2024). Since this effect is in the opposite direction of what we see, we do not consider this scenario further.

## Evidence for heterozygous effects.

We next considered an alternative model in which pathogenic variants within recessive disease genes have effects in heterozygotes that are nonclinical, yet large enough to be relevant for selection. Since heterozygotes are so much more common than homozygotes, even weak selection in heterozygotes may be relevant for determining  $p_{\text{path}}$  (Amorim et al. 2017).

Prior work provides evidence that stabilizing selection may be the primary mode of selection for quantitative traits (Simons et al. 2018; Sella and Barton 2019), including for LoFs (Milind et al. 2024; Spence et al. 2025). In this model, the selection against LoF heterozygotes ( $s_{\text{het}}$ ), or other genetic variation, is proportional to a sum of squared effect sizes of that variation on all regulated traits (Fig. 3a) (Simons et al. 2018; Sanjak et al. 2018; Sella and Barton 2019; Simons et al. 2025).

How might this model be relevant to recessive genes? To conceptualize this, we consider the molecular effects of variants that change expression of a relevant gene between 0% (e.g. homozygous LoFs) to 50% (e.g. heterozygous LoFs) to 100% of typical expression. We model these different expression levels as corresponding to different expected values in one of the trait dimensions (Fig. 3b) (Naqvi et al. 2023; Milind et al. 2024; Domingo et al. 2025). We propose that, for recessive disease genes, a single LoF may impact quantitative traits, while being insufficient to cause a clinical disease state (Fig. 3b) (Fridman et al. 2021; Barton et al. 2022; Zschocke et al. 2023; Heyne et al. 2023; Fridman et al. 2025).

To test this model we sought to measure the effects of heterozygous LoFs at recessive genes on quantitative traits measured by the UKB. We performed LoF burden tests on 19,194 genes, including our curated list of 796 recessive disease genes, for 161 quantitative traits. To ensure that we specifically capture the role of heterozygous LoFs on these quantitative traits, we excluded homozygous LoF individuals from the data analysis.

Our group has previously shown that, considering all genes, there is a strong relationship between selective constraint and the average squared effect size of LoFs on quantitative traits (Spence et al. 2025).

For nonrecessive genes, heterozygous gene constraint is estimated using GeneBayes, a Bayesian estimator that provides well-calibrated  $s_{\text{het}}$  estimates regardless of observed  $p_{\text{LoF}}$  (Zeng et al. 2024); for recessive genes,  $s_{\text{het}}$  is instead estimated as a lower bound derived from additive mutation–selection balance applied to the observed  $p_{\text{LoF}}$ . Because these two constraint measures are derived through methodologically distinct pipelines, the correspondence in burden test enrichment across deciles serves as a consistency check on the core hypothesis: if stabilizing selection on heterozygous trait effects drives constraint in recessive genes, then the mutation–selection balance lower bound on  $s_{\text{het}}$  should predict burden enrichment in a manner qualitatively similar to GeneBayes-estimated constraint in nonrecessive genes, as we observe (Fig. 3c).

Were pleiotropic heterozygous effects negligible for recessive genes, the recessive curve would be expected to remain flat across all deciles regardless of constraint level; the close correspondence with the nonrecessive curve instead supports the interpretation that the additive  $s_{\text{het}}$  estimates for recessive genes reflects genuine heterozygous fitness consequences. Additionally, a sensitivity analysis in which estimated  $s_{\text{het}}$  are scaled downward by up to 75% to account for potential confounding by a combination of inbreeding, drift, and underestimation of  $p_{\text{LoF}}$  indicates that the qualitative pattern is robust across this range of assumptions (Supplementary Fig. 5).

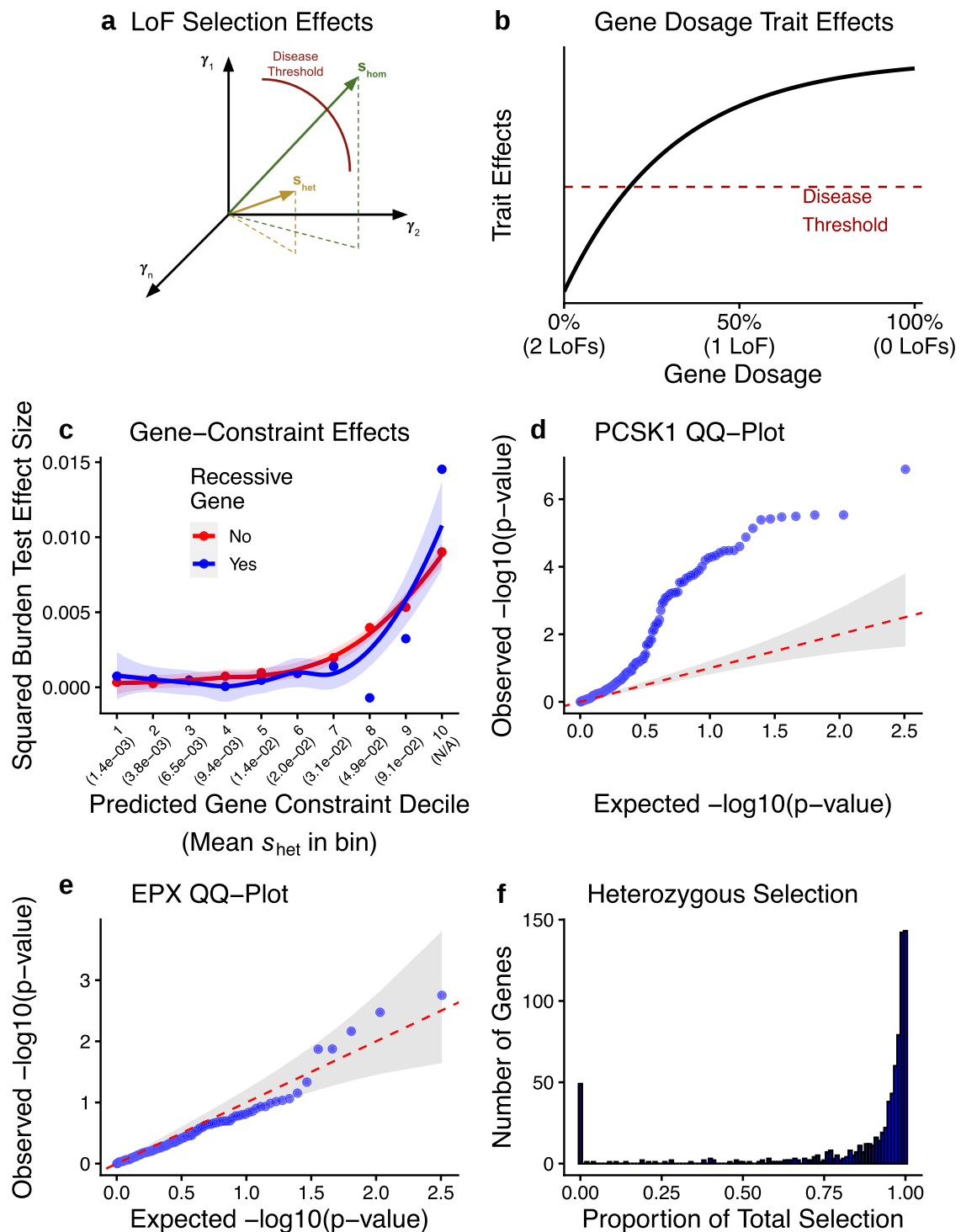
In summary, this analysis suggests that widespread heterozygous effects on quantitative traits in recessive disease genes may drive the globally poor fit of the recessive mutation–selection balance model.

**Significance testing of gene–trait pairs.** We also tested the effect of the recessive genes on each of 161 quantitative traits. Although these tests are generally underpowered due to low LoF frequencies and mostly modest effect sizes, we find that 142 of 796 (17.8%) recessive disease genes have significant effects on one or more traits in heterozygotes (Aggregated Cauchy Association Test (ACAT)  $p$ -value  $< 0.05$ ) (Supplementary Fig. 6a). We find qualitatively similar results when we restrict this analysis to 21 uncorrelated quantitative traits (11.8% of traits significant; Supplementary Fig. 6b). These percentages are similar to what we find for the heterozygous effects of all nonrecessive genes using the same LoF burden test results (14.6 % for all traits; 14.0 % for uncorrelated traits) (Supplementary Fig. 6c and d).

One example of a recessive disease gene with significant heterozygous effects is PCSK1, a gene in the same family as PCSK9 (known for its effects on LDL cholesterol levels (Cariou et al. 2011)). While PCSK1 is known to cause endocrinopathy in an autosomal recessive fashion (Cool et al. 1997; Jackson et al. 2003), we find that it has effects on BMI, fat accumulation, and other weight-related phenotypes in heterozygotes (Fig. 3d, Supplementary Fig. 6e, and Supplementary Table 5). This indicates that while homozygous LoF of PCSK1 results in clinical endocrinopathy, variation in PCSK1 expression such as through heterozygous LoFs have effects on quantitative traits, potentially resulting in selection against such variants (Ramos-Molina et al. 2016; Stijnen et al. 2014, 2016).

In contrast, the recessive gene EPX has no observed heterozygous effects and the observed LoF frequency fits recessive mutation–selection balance. EPX plays a role in tyrosine nitration granule proteins in resting eosinophils and is recessively linked to eosinophil peroxidase deficiency (Fig. 3e and Supplementary Fig. 6f) (Percopo et al. 2019). While extensive research has identified an association between EPX and various phenotypes, such as asthma, allergic response, and airway function, these relationships have primarily been identified in studies with total gene knockouts (Nair et al. 2013; Klonoff-Cohen and Polavarapu 2016; Percopo et al. 2019; Acharya and Ackerman 2014). However, GWAS have identified associations between EPX and molecular traits, such as eosinophil and basophil count (Barton et al. 2021), suggesting that even this gene likely has some degree of phenotypic effects at small-effect variants.

**Estimates of the fraction of heterozygous selection.** Lastly, the estimator for  $s_{\text{het}}$  derived from additive mutation–selection balance assuming  $s_{\text{hom}} = 1$  also permits a direct calculation of the minimum fraction of selective pressure acting through heterozygotes for each recessive gene. We sought to estimate, for each recessive gene, the minimum fraction of selective removal of pathogenic



**Fig. 3.** Heterozygous burden effects for recessive genes: a) variant effects can be conceptualized as vectors in a high-dimensional trait space (Simons et al. 2018), shown here for hypothetical heterozygous and homozygous LoFs. Here, only the homozygous LoF crosses a clinical threshold for disease, but both genotypes are subject to selection (selection is proportional to the squared magnitude of the vector). b) Hypothetical relationship between gene expression dosage and model trait effects (Naqvi et al. 2023). The red line represents the trait effect level where a clinical diagnosis would be made. c) Relationship between estimated heterozygous gene constraint decile and average bias-corrected squared LoF burden test effect sizes for 161 quantitative traits, split according to whether or not the genes are classified as recessive. Parentheses denote the mean predicted  $s_{het}$  within each decile bin, calculated across both recessive and nonrecessive genes. The tenth decile is labeled N/A because  $s_{het}$  is indeterminate for recessive genes with  $p_{LoF} = 0$  under the recessive-gene constraint estimation method. Lines show loess curve fits and standard errors describing the relationship between constraint and squared burden effect sizes. d) Quantile–quantile plot of 161 LoF burden test  $p$ -values for the recessive PCSK1 gene. e) Quantile–quantile plot of 161 LoF burden test  $p$ -values for the recessive EPX gene. f) Histogram showing the numbers of recessive genes with a given proportion of total selection acting through heterozygous LoFs ( $f_{het}$ ).

mutations that occurs in heterozygotes, which we refer to as  $f_{\text{het}}$ , and define as

$$f_{\text{het}} = \frac{2p_{\text{path}}(1 - p_{\text{path}})s_{\text{het}}}{2p_{\text{path}}(1 - p_{\text{path}})s_{\text{het}} + 2p_{\text{path}}^2s_{\text{hom}}}. \quad (9)$$

Since  $f_{\text{het}}$  depends on the unknown values of  $s_{\text{het}}$  and  $s_{\text{hom}}$ , we set

$$\widehat{f_{\text{het}}} = 0, \quad \text{if } p_{\text{path}} \geq \sqrt{\mu L}, \quad (10)$$

which conservatively assumes that the allele frequencies are consistent with fully recessive selection; and otherwise we compute a lower bound on  $f_{\text{het}}$  as follows:

$$\widehat{f_{\text{het}}} \leq 1 - \frac{p_{\text{path}}^2}{\mu L}, \quad (11)$$

which makes the conservative assumption that homozygotes have zero fitness, i.e.  $s_{\text{hom}} = 1$ , to maximize the amount of selection occurring in homozygotes (Supplementary Appendix A).

Applying this to empirical data, we find that for most genes, most selection acts through heterozygotes, even when we assume LoFs are homozygous lethal. Out of all recessive genes, only 74 genes (9%) are consistent with being primarily driven by homozygous selection (Fig. 3f). Under the conservative assumption of recessive lethality, we estimate that for 73% of annotated recessive genes, at least 90% of the selection acts through heterozygotes; these estimates represent lower bounds on  $s_{\text{het}}$  and are therefore not contingent on precise knowledge of  $s_{\text{hom}}$ .

## High-frequency recessive genes

Although most recessive disease genes do not appear to follow recessive mutation–selection balance, a handful of genes in each population do have predicted pathogenic variation frequencies within the range consistent with the fully recessive model (Fig. 4a). While the specific sets of high-frequency genes vary (Fig. 4b–d and Supplementary Table 6), the absolute numbers are quite similar across populations (49 in NFE, 34 in FIN, 39 in AFR, and 43 in SAS).

Genetic drift, and potential variant ascertainment bias (Stolyarova et al. 2025) may explain the differences between populations, and this stochasticity may result in some genes having higher than expected  $p_{\text{path}}$ . Alternatively, the difference in which genes are consistent with no selection against heterozygotes may be driven by different selective pressures in different populations, for example due to differing environments. Finally, here we only considered selection against LoFs, but these genes with high  $p_{\text{path}}$  could also be driven by balancing selection or positive selection (Andrés et al. 2009; Booker et al. 2017).

When we examined in different populations genes previously believed to be affected by balancing selection (Clark 1998; Risch et al. 2003; Andrés et al. 2009; Alfonso-Sánchez et al. 2010; Garbade et al. 2019), we observed variation in the genes that fit recessive mutation–selection balance (Supplementary Fig. 7a–d). For example, two recessive disease genes that fit mutation–selection balance based on the study population and are believed to be affected by balancing selection are PAH (Supplementary Fig. 7a) and CFTR (Supplementary Fig. 7d), which are the genes associated with causing PKU (Oussalah et al. 2020) and cystic fibrosis (Alfonso-Sánchez et al. 2010), respectively. If these genes and other recessive disease genes are under the influence of balancing selection, their  $p_{\text{path}}$  would be higher than expected under a model

of purely recessive selection, and potentially look consistent with recessive mutation–selection balance based on the specific population analyzed, due to differences in selective pressures, and choice of variant or LoF inclusion criteria.

Alternatively, this analysis may identify genes where the selection pressures acting on the gene and affecting  $p_{\text{path}}$  are unknown. A potential example among these recessive genes is PCK2, a mitochondrial isozyme of PEPCK, which has a well-known function in gluconeogenesis (Duan 2024). PCK2 is consistent with having no selection against heterozygotes in both the NFE and SAS populations (Fig. 4a and d), and possesses particularly high  $p_{\text{path}}$  in the NFE population (Fig. 4a and Supplementary Table 2). When examining the phenotypic effects of this gene, we observe a wide range of heterozygous effects on the 161 traits we analyzed (Fig. 4e), and other research has found that over-expression of PCK2 is linked to lung cancer risk, diabetes, and obesity (Xue et al. 2024; Vincent et al. 2015). This suggests that perhaps any selection against PCK2 in heterozygotes caused by these phenotypic consequences is offset by some other, unknown fitness advantage.

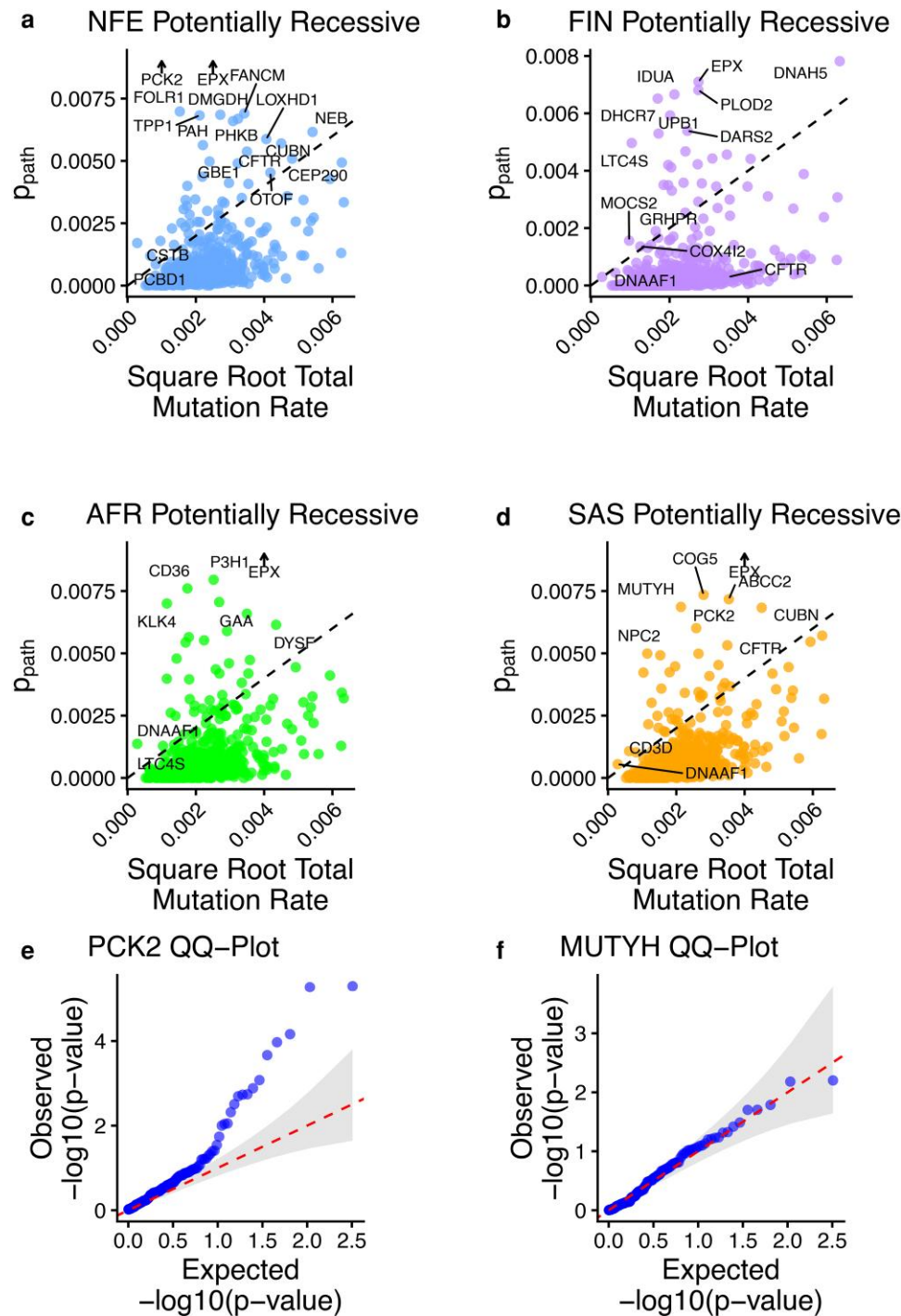
All together, these results indicate that even these genes that are consistent with purely recessive selection may have effects in heterozygotes, and need not follow the textbook definition of recessive selection.

On the other hand, some genes appear to be purely recessive by all measures we examined. For example, *MUTYH* fits recessive mutation–selection balance (Fig. 4d) and has no observed heterozygous effects (Fig. 4f). *MUTYH* has a role in base excision repair and is recessively associated with colorectal adenoma risk (Hemker et al. 2025). Recent research has found that *MUTYH* continues to function in suppressing mutations even in the heterozygous LoF state (Young et al. 2023), and only when two LoFs are present does the rate of *de novo* mutations increase, resulting in increased cancer risk. Overall, it seems plausible that *MUTYH* is an example of a gene governed by recessive mutation–selection balance.

We also examined the subset of recessive genes that match the predictions of purely recessive mutation–selection balance in the NFE, FIN, AFR, and SAS populations (Fig. 4a–d) to identify whether these genes showed any particular patterns. The associated diseases as well as cellular functions of these genes were varied and not limited to a specific biological function or system. However, genes that fit mutation–selection balance in the NFE population are enriched compared to all recessive disease genes for polarized epithelial cellular component GO terms (Supplementary Table 7,  $p$ -value < 0.05), which are involved in transport and environmental sensing and have been associated with a variety of diseases including immunological effects (Wilson 1997; Schleimer and Berdnikovs 2017; Lu et al. 2024), possibly indicating positive or balancing selection driven by host/pathogen competition (von Boehmer 1994; Hedrick 2007; Antonovics et al. 2013; Sironi et al. 2015).

## Discussion

The role of heterozygous and homozygous selection on the equilibrium allele frequency of pathogenic variation in a gene has been understood theoretically for nearly a century (Haldane 1937). Purely recessive mutation–selection balance is a classic result found in textbooks and often taught in introductory population genetics courses (Lewontin 1985; Crow 2017). Meanwhile, the study of Mendelian autosomal recessive genes might lead one to believe that many genes should be governed by purely recessive mutation–selection balance.



**Fig. 4.** Recessive genes under potentially recessive selection: observed relationship between square root total mutation rate and predicted  $p_{\text{path}}$  for recessive disease genes. Chosen recessive disease genes with a predicted  $p_{\text{path}}$  in the expected range for recessive mutation–selection balance are labeled using a) NFE allele frequencies, b) FIN European allele frequencies, c) AFR population allele frequencies, and d) SAS population allele frequencies. e) Quantile–quantile plot of 161 LoF burden test  $p$ -values for the recessive PCK2 gene. f) Quantile–quantile plot of 161 LoF burden test  $p$ -values for the recessive MUTYH gene.

Here, we find purely recessive mutation–selection balance is a poor fit for nearly every protein coding gene in the human genome, including recessive disease genes. Our analyses reveal that inbreeding is an unlikely explanation for this phenomenon, and observed  $p_{\text{LoF}}$  are instead likely shaped by selection against heterozygotes. We emphasize that the departure from the purely recessive model does not reflect a fundamental limitation of the mutation–selection balance framework, which has been shown

to provide reasonable approximations of allele frequencies for genes with well-characterized selective effects (Cassa et al. 2017; Weghorn et al. 2019). Rather, we find that heterozygous LoFs in recessive disease genes have quantitative effects that are similar to genes overall, consistent with prior work demonstrating that isolating recessive selection coefficients from population frequency data is inherently challenging in the presence of substantial heterozygous effects (Balick et al. 2022).

While these effects might not rise to the level of being classified as Mendelian disorders, the substantially larger frequency of individuals with heterozygous LoFs means that even modest selection against heterozygotes may contribute substantially to determining  $p_{\text{LoF}}$  in recessive disease genes, and for a meaningful proportion of annotated recessive genes our analyses indicate that heterozygous selection is the dominant force. Our results suggest that recessive disease genes are often highly pleiotropic, with effects even in heterozygotes (Barton et al. 2022; Heyne et al. 2023; Fridman et al. 2025), and hence could play roles in many complex diseases. One possible explanation could be that stabilizing selection on a multidimensional phenotypic space naturally generates partial recessivity for fitness even when mutations act additively on the underlying traits (Manna et al. 2011). As a result, any model of selection in humans should default to assuming that selection occurring in heterozygotes is a major factor.

The broad applicability of purely recessive mutation–selection balance had already been questioned in the pregenomics era, at least in *Drosophila*. By 1941, Dobzhansky and Wright (and later in work with Hovanitz) found that recessive lethal chromosomes in *Drosophila pseudoobscura* were too rare to be explained by recessive mutation selection balance and posited that either selection in heterozygotes or inbreeding were driving the frequencies to be lower (Dobzhansky and Wright 1941; Wright et al. 1942). In 1968, Nei re-examined this question by using the mean and variance of the frequency of recessive lethal chromosomes across different *Drosophila* populations, and came to a similar conclusion (Nei 1968). As summarized in Charlesworth and Charlesworth (2010), later work came to broadly similar conclusions but discounted the effects of inbreeding by estimating heterozygous fitness effects of recessive lethal alleles (Crow 1993; García-Dorado and Caballero 2000).

Another aspect of our study worth considering is that the traits tested for heterozygous effects through burden tests are relatively limited. While these traits were chosen to represent a range of continuous nondisease, biochemistry, and anthropometric traits, they are far from exhaustive. Thus, our results provide a conservative estimate of the role of heterozygous pathogenic variants in recessive disease genes, and inclusion of more uncorrelated traits, particularly molecular traits, should only reveal an even broader range of heterozygous effects of variants in recessive genes (Burton et al. 2009; Smith et al. 2022; Ota et al. 2025).

Furthermore, it is difficult to determine if genes with heterozygous effects are classified as recessive because they are understudied, or if the heterozygous effects are too small to be clinically relevant. For example, recessive genes such as PCSK1 have strong heterozygous effects that are easily identified through LoF burden tests. While identifying these effects can inform our understanding of a wide range of quantitative traits, classifying these strong effect genes as recessive may simply be due to a lack of research. If these genes were excluded, the mutation–selection balance model may appear more valid in its current form. Inversely, there are a large number of recessive genes where we infer that the majority of selection acts through heterozygotes, but no significant gene–trait effects were identified. This can be partially attributed to the previously mentioned issue of a lack of tested traits or a lack of power, insofar as current datasets may be too small to identify multiple subthreshold effects across traits not captured in the current analysis, which account for the effect of variations in these genes over many generations. Resolving this more conclusively will likely require burden testing against larger panels of molecular and cellular phenotypes, as well as increased sample sizes that provide greater power to

detect the modest per-trait effect sizes expected under a model of pleiotropic stabilizing selection. Identifying this second class of genes with known recessive effects and minuscule additive effects is particularly crucial for understanding human genetic variation.

In summary, we have explored the relationship between equilibrium allele frequency and natural selection on complex traits. Our results highlight the potential for widespread effects of heterozygous LoFs, even in recessive disease genes. These insights about recessive genes and natural selection will be informative for various downstream analyses, including the study of the genetic architecture of complex traits and investigations of the selective pressures underlying human genetic variation.

## Data availability

UK Biobank data are publicly available by request from <https://www.ukbiobank.ac.uk>. The research was conducted with approved access to UK Biobank data under application number 52374 (PI: pritchard).

Supplemental material available at [GENETICS](https://www.genetics.org/genetics-advance-article/doi/10.1093/genetics/iyag150/8707824) online.

## Acknowledgments

We thank Josh Schraiber for helpful discussions, Molly Przeworski for helpful comments on the manuscript, and Shai Carmi, Graham Coop, and Hilary Martin for pointing us to related work. We additionally thank our anonymous reviewers for their comments that have helped to improve this work. Anastasia Stolyarova kindly provided processed data from another study (Stolyarova et al. 2025). Thanks to the Pritchard Lab for many valuable comments and discussions.

## Funding

This work was supported by the National Institutes of Health grants R01HG014005, R01HG008140 (to JKP) and R01HL175076 (JPS).

## Conflicts of interest

None declared.

## Literature cited

- Acharya KR, Ackerman SJ. 2014. Eosinophil granule proteins: form and function. *J Biol Chem*. 289:17406–17415. <https://doi.org/10.1074/jbc.R113.546218>.
- Adam MP et al., editors. 1993. GeneReviews®. University of Washington.
- Agarwal I, Fuller ZL, Myers SR, Przeworski M. 2023. Relating pathogenic loss-of-function mutations in humans to their evolutionary fitness costs. *Elife*. 12:e83172. <https://doi.org/10.7554/eLife.83172>.
- Alfonso-Sánchez MA, Pérez-Miranda AM, García-Obregón S, Peña JA. 2010. An evolutionary approach to the high frequency of the Delta F508 CFTR mutation in European populations. *Med Hypotheses*. 74:989–992. <https://doi.org/10.1016/j.mehy.2009.12.018>.
- Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: online Mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 43:D789–D798. <https://doi.org/10.1093/nar/gku1205>.

- Amorim CEG et al. 2017. The population genetics of human disease: the case of recessive, lethal mutations. *PLoS Genet.* 13: e1006915. <https://doi.org/10.1371/journal.pgen.1006915>.
- Andrés AM et al. 2009. Targets of balancing selection in the human genome. *Mol Biol Evol.* 26:2755–2764. <https://doi.org/10.1093/molbev/msp190>.
- Antonovics J et al. 2013. The origin of specificity by means of natural selection: evolved and nonhost resistance in host–pathogen interactions. *Evolution.* 67:1–9. <https://doi.org/10.1111/j.1558-5646.2012.01793.x>.
- Ashburner M et al. 2000. Gene ontology: tool for the unification of biology. *Nat Genet.* 25:25–29. <https://doi.org/10.1038/75556>.
- Balick DJ, Jordan DM, Sunyaev S, Do R. 2022. Overcoming constraints on the detection of recessive selection in human genes from population frequency data. *Am J Hum Genet.* 109:33–49. <https://doi.org/10.1016/j.ajhg.2021.12.001>.
- Barton AR, Hujoel ML, Mukamel RE, Sherman MA, Loh P-R. 2022. A spectrum of recessiveness among Mendelian disease variants in UK Biobank. *Am J Hum Genet.* 109:1298–1307. <https://doi.org/10.1016/j.ajhg.2022.05.008>.
- Barton AR, Sherman MA, Mukamel RE, Loh P-R. 2021. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat Genet.* 53:1260–1269. <https://doi.org/10.1038/s41588-021-00892-1>.
- Berg JS et al. 2013. An informatics approach to analyzing the incidenta-lome. *Genet Med.* 15:36–44. <https://doi.org/10.1038/gim.2012.112>.
- Blekhman R et al. 2008. Natural selection on genes that underlie human disease susceptibility. *Curr Biol.* 18:883–889. <https://doi.org/10.1016/j.cub.2008.04.074>.
- Booker TR, Jackson BC, Keightley PD. 2017. Detecting positive selection in the genome. *BMC Biol.* 15:98. <https://doi.org/10.1186/s12915-017-0434-y>.
- Boyle EA, Li YI, Pritchard JK. 2017. An expanded view of complex traits: from polygenic to omnigenic. *Cell.* 169:1177–1186. <https://doi.org/10.1016/j.cell.2017.05.038>.
- Burton PR et al. 2009. Size matters: just how big is BIG?: quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol.* 38:263–273. <https://doi.org/10.1093/ije/dyn147>.
- Bycroft C et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 562:203–209. <https://doi.org/10.1038/s41586-018-0579-z>.
- Cariou B, Le May C, Costet P. 2011. Clinical aspects of PCSK9. *Atherosclerosis.* 216:258–265. <https://doi.org/10.1016/j.atherosclerosis.2011.04.018>.
- Cassa CA et al. 2017. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat Genet.* 49:806–810. <https://doi.org/10.1038/ng.3831>.
- Castellani C. 2013. CFTR2: how will it help care? *Paediatr Respir Rev.* 14:2–5. <https://doi.org/10.1016/j.prrv.2013.01.006>.
- Castiglione A, Möller C. 2022. Usher syndrome. *Audiol Res.* 12:42–65. <https://doi.org/10.3390/audiolres12010005>.
- Charlesworth B, Charlesworth D. 2010. Elements of evolutionary genetics.
- Charlesworth D, Willis JH. 2009. The genetics of inbreeding depression. *Nat Rev Genet.* 10:783–796. <https://doi.org/10.1038/nrg2664>.
- Chen S et al. 2024. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature.* 625:92–100. <https://doi.org/10.1038/s41586-023-06045-0>.
- Clark AG. 1998. Mutation-selection balance with multiple alleles. In: Woodruff RC, Thompson JN, editors. *Mutation and evolution. Contemporary issues in genetics and evolution*, Vol. 7. Springer. p. 41–47. [https://doi.org/10.1007/978-94-011-5210-5\\_4](https://doi.org/10.1007/978-94-011-5210-5_4).
- Cool DR et al. 1997. Carboxypeptidase E is a regulated secretory pathway sorting receptor: genetic obliteration leads to endocrine disorders in *Cpefat* mice. *Cell.* 88:73–83. [https://doi.org/10.1016/S0092-8674\(00\)81860-7](https://doi.org/10.1016/S0092-8674(00)81860-7).
- Crow JF. 1993. How much do we know about spontaneous human mutation rates? *Environ Mol Mutagen.* 21:122–129. <https://doi.org/10.1002/em.2850210205>.
- Crow JF. 2017 Jan 1. *An introduction to population genetics theory*. Scientific Publishers. 609 pp.
- Cummings BB et al. 2020. Transcript expression-aware annotation improves rare variant interpretation. *Nature.* 581:452–458. <https://doi.org/10.1038/s41586-020-2329-2>.
- Dobzhansky T, Wright S. 1941. *Genetics of natural populations. V. Relations between mutation rate and accumulation of lethals in populations of drosophila pseudoobscura*. Genetics. 26:23. <https://doi.org/10.1093/genetics/26.1.23>.
- Domingo J et al. 2025. Nonlinear transcriptional responses to gradual modulation of transcription factor dosage. <https://doi.org/10.1101/2024.03.01.582837>.
- Duan KL et al. 2024. PCK2 maintains intestinal homeostasis and prevents colitis by protecting antibody-secreting cells from oxidative stress. *Immunology.* 173:339–359. <https://onlinelibrary.wiley.com/doi/10.1111/imm.13827>.
- Durbin RM et al. 2010. A map of human genome variation from population-scale sequencing. *Nature.* 467:1061–1073. <https://doi.org/10.1038/nature09534>.
- Fernandes Filho JA, Shapiro BE. 2004. Tay–Sachs disease. *Arch Neurol.* 61:1466–1468. <https://doi.org/10.1001/archneur.61.9.1466>.
- Fridman H et al. 2021. The landscape of autosomal-recessive pathogenic variants in European populations reveals phenotype-specific effects. *Am J Hum Genet.* 108:608–619. <https://doi.org/10.1016/j.ajhg.2021.03.004>.
- Fridman H, Khazeeva G, Levy-Lahad E, Gilissen C, Brunner HG. 2025. Reproductive and cognitive phenotypes in carriers of recessive pathogenic variants. *Nat Hum Behav.* 9:1726–1736. <https://doi.org/10.1038/s41562-025-02204-7>.
- Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. 2019. Measuring intolerance to mutation in human genetics. *Nat Genet.* 51:772–776. <https://doi.org/10.1038/s41588-019-0383-1>.
- Garbade SF et al. 2019. Allelic phenotype values: a model for genotype-based phenotype prediction in phenylketonuria. *Genet Med.* 21:580–590. <https://doi.org/10.1038/s41436-018-0081-x>.
- García-Dorado A, Caballero A. 2000. On the average coefficient of dominance of deleterious spontaneous mutations. *Genetics.* 155:1991–2001. <https://doi.org/10.1093/genetics/155.4.1991>.
- Gazal S, Sahbatou M, Babron M-C, Génin E, Leutenegger A-L. 2015. High level of inbreeding in final phase of 1000 genomes project. *Sci Rep.* 5:17453. <https://doi.org/10.1038/srep17453>.
- Guo MH et al. 2024. Inferring compound heterozygosity from large-scale exome sequencing data. *Nat Genet.* 56:152–161. <https://doi.org/10.1038/s41588-023-01608-3>.
- Haldane JBS. 1927. A mathematical theory of natural and artificial selection, part V: selection and mutation. *Math Proc Camb Philos Soc.* 23:838–844. <https://doi.org/10.1017/S0305004100015644>.
- Haldane JBS. 1937. The effect of variation of fitness. *Am Nat.* 71:337–349. <https://doi.org/10.1086/280722>.
- Haller BC, Messer PW. 2023. SLiM 4: multispecies eco-evolutionary modeling. *Am Nat.* 201:E127–E139. <https://doi.org/10.1086/723601>.
- Hedrick PW. 2007. Balancing selection. *Curr Biol.* 17:R230–R231. <https://doi.org/10.1016/j.cub.2007.01.012>.
- Hedrick PW, García-Dorado A. 2016. Understanding inbreeding depression, purging, and genetic rescue. *Trends Ecol Evol.* 31:940–952. <https://doi.org/10.1016/j.tree.2016.09.005>.

- Hemker SL et al. 2025. Saturation mapping of MUTYH variant effects using DNA repair reporters. *Am J Hum Genet.* 112:2010–2026. <https://doi.org/10.1016/j.ajhg.2025.07.005>.
- Heyne H et al. 2023. Mono-and biallelic variant effects on disease at biobank scale. *Nature.* 613:519–525. <https://doi.org/10.1038/s41586-022-05420-7>.
- Jackson M, Marks L, May GH, Wilson J. 2018. The genetic basis of disease. *Essays Biochem.* 62:643–723. <https://doi.org/10.1042/EBC20170053>.
- Jackson RS et al. 2003. Small-intestinal dysfunction accompanies the complex endocrinopathy of human proprotein convertase 1 deficiency. *J Clin Invest.* 112:1550–1560. <https://doi.org/10.1172/JCI18784>.
- Karczewski KJ et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 581:434–443. <https://doi.org/10.1038/s41586-020-2308-7>.
- Klonoff-Cohen H, Polavarapu M. 2016. Eosinophil protein X and childhood asthma: a systematic review and meta-analysis. *Immun Inflamm Dis.* 4:114–134. <https://doi.org/10.1002/iid3.104>.
- Kumar M et al. 2020. WilsonGen a comprehensive clinically annotated genomic variant resource for wilson's disease. *Sci Rep.* 10: 9037. <https://doi.org/10.1038/s41598-020-66099-2>.
- Kundaje A et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature.* 518:317–330. <https://doi.org/10.1038/nature14248>.
- Landrum M et al. 2025. ClinVar: updates to support classifications of both germline and somatic variants. *Nucleic Acids Res.* 53: D1313–D1321. <https://doi.org/10.1093/nar/gkae1090>.
- Lassen FH et al. 2024. Exome-wide evidence of compound heterozygous effects across common phenotypes in the UK Biobank. *Cell Genom.* 4:100602. <https://doi.org/10.1016/j.xgen.2024.100602>.
- Laws RJ, Jamieson IG. 2011. Is lack of evidence of inbreeding depression in a threatened New Zealand robin indicative of reduced genetic load?. *Anim Conserv.* 14:47–55. <https://doi.org/10.1111/j.1469-1795.2010.00388.x>.
- Lek M et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 536:285–291. <https://doi.org/10.1038/nature19057>.
- Lewontin RC. 1985. Population genetics. *Annu Rev Genet.* 19:81–102. <https://doi.org/10.1146/annurev.ge.19.120185.000501>.
- Liu Y et al. 2019. ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am J Hum Genet.* 104:410–421. <https://doi.org/10.1016/j.ajhg.2019.01.002>.
- Loneragan GJ, Cline DB, Abbondanzo SL. 2001. Sickle cell anemia. *Radiographics.* 21:971–994. <https://doi.org/10.1148/radiographics.21.4.g01j123971>.
- Lu H-F et al. 2024. Involvement and repair of epithelial barrier dysfunction in allergic diseases. *Front Immunol.* 15:1348272. <https://doi.org/10.3389/fimmu.2024.1348272>.
- Manna F, Martin G, Lenormand T. 2011. Fitness landscapes: an alternative theory for the dominance of mutation. *Genetics.* 189: 923–937. <https://doi.org/10.1534/genetics.111.132944>.
- Mbatchou J et al. 2021. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet.* 53: 1097–1103. <https://doi.org/10.1038/s41588-021-00870-7>.
- McKusick VA. 2024. Mendelian inheritance in man: catalogs of autosomal dominant, autosomal recessive, and x-linked phenotypes. Elsevier. 368 pp.
- McLaren W et al. 2016. The ensembl variant effect predictor. *Genome Biol.* 17:122. <https://doi.org/10.1186/s13059-016-0974-4>.
- Mercuri E, Sumner CJ, Muntoni F, Darras BT, Finkel RS. 2022. Spinal muscular atrophy. *Nat Rev Dis Primers.* 8:52. <https://doi.org/10.1038/s41572-022-00380-8>.
- Milind N, Smith CJ, Zhu H, Spence JP, Pritchard JK. 2024 Nov 11. Buffering and non-monotonic behavior of gene dosage response curves for human complex traits. <https://doi.org/10.1101/2024.11.11.24317065v1>.
- Molotkov I, Mardis ER, Artomov M. 2024. Making sense of missense: challenges and opportunities in variant pathogenicity prediction. *Dis Model Mech.* 17:dmm052218. <https://doi.org/10.1242/dmm.052218>.
- Nagral A. 2014. Gaucher disease. *J Clin Exp Hepatol.* 4:37–50. <https://doi.org/10.1016/j.jceh.2014.02.005>.
- Nair P et al. 2013. Eosinophil peroxidase in sputum represents a unique biomarker of airway eosinophilia. *Allergy.* 68:1177–1184. <https://doi.org/10.1111/all.12206>.
- Naqvi S et al. 2023. Precise modulation of transcription factor levels identifies features underlying dosage sensitivity. *Nat Genet.* 55: 841–851. <https://doi.org/10.1038/s41588-023-01366-2>.
- Nei M. 1968. The frequency distribution of lethal chromosomes in finite populations. *Proc Natl Acad Sci U S A.* 60:517–524. <https://doi.org/10.1073/pnas.60.2.517>.
- Ong T, Ramsey BW. 2023. Cystic fibrosis: a review. *JAMA.* 329: 1859–1871. <https://doi.org/10.1001/jama.2023.8120>.
- Ota M et al. 2025. Causal modeling of gene effects from regulators to programs to traits: integration of genetic associations and Perturb-seq. <https://doi.org/10.1101/2025.01.22.634424>.
- Oussalah A et al. 2020. Population and evolutionary genetics of the PAH locus to uncover overdominance and adaptive mechanisms in phenylketonuria: results from a multiethnic study. *EBioMedicine.* 51:102623. <https://doi.org/10.1016/j.ebiom.2019.102623>.
- Percopo CM et al. 2019. Impact of eosinophil-peroxidase (EPX) deficiency on eosinophil structure and function in mouse airways. *J Leukoc Biol.* 105:151–161. <https://doi.org/10.1002/jlb.3AB0318-090RR>.
- Ramos-Molina B, Martin MG, Lindberg I. 2016. PCSK1 variants and human obesity. In: Tao Y-X, editor. *Progress in molecular biology and translational science, Genetics of monogenic and syndromic obesity*, Vol. 140. Academic Press. 47–74. <https://doi.org/10.1016/bs.pmbts.2015.12.001>.
- Risch N, Tang H, Katzenstein H, Ekstein J. 2003. Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. *Am J Hum Genet.* 72: 812–822. <https://doi.org/10.1086/373882>.
- Robinson JA, Brown C, Kim BY, Lohmueller KE, Wayne RK. 2018. Purging of strongly deleterious mutations explains long-term persistence and absence of inbreeding depression in island foxes. *Curr Biol.* 28:3487–3494.e4. <https://doi.org/10.1016/j.cub.2018.08.066>.
- Sanjak JS, Sidorenko J, Robinson MR, Thornton KR, Visscher PM. 2018. Evidence of directional and stabilizing selection in contemporary humans. *Proc Natl Acad Sci U S A.* 115:151–156. <https://doi.org/10.1073/pnas.1707227114>.
- Schleimer RP, Berdnikovs S. 2017. Etiology of epithelial barrier dysfunction in patients with type 2 inflammatory diseases. *J Allergy Clin Immunol.* 139:1752–1761. <https://doi.org/10.1016/j.jaci.2017.04.010>.
- Seal RL et al. 2023. [Genenames.org](https://www.genenames.org): the HGNC resources in 2023. *Nucleic Acids Res.* 51:D1003–D1009. <https://doi.org/10.1093/nar/gkac888>.
- Sella G, Barton NH. 2019. Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annu Rev Genomics Hum Genet.* 20:461–493. <https://doi.org/10.1146/annurev-genom-083115-022316>.
- Simons YB et al. 2025. Simple scaling laws control the genetic architectures of human complex traits. *PLoS Biol.* 23:e3003402. <https://doi.org/10.1371/journal.pbio.3003402>.

- Simons YB, Bullaughey K, Hudson RR, Sella G. 2018. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biol.* 16:e2002985. <https://doi.org/10.1371/journal.pbio.2002985>.
- Sironi M, Cagliani R, Forni D, Clerici M. 2015. Evolutionary insights into host–pathogen interactions from mammalian sequence data. *Nat Rev Genet.* 16:224–236. <https://doi.org/10.1038/nrg3905>.
- Smith CJ et al. 2022. Integrative analysis of metabolite GWAS illuminates the molecular basis of pleiotropy and genetic correlation. *Life.* 11:e79348. <https://doi.org/10.7554/eLife.79348>.
- Spence JP et al. 2025. Specificity, length and luck drive gene rankings in association studies. *Nature.* 649:918–925. <https://doi.org/10.1038/s41586-025-09703-7>.
- Stijnen P et al. 2014. The association of common variants in PCSK1 with obesity: a HuGE review and meta-analysis. *Am J Epidemiol.* 180:1051–1065. <https://doi.org/10.1093/aje/kwu237>.
- Stijnen P, Ramos-Molina B, O’Rahilly S, Creemers JWM. 2016. PCSK1 mutations and human endocrinopathies: from obesity to gastrointestinal disorders. *Endocr Rev.* 37:347–371. <https://doi.org/10.1210/er.2015-1117>.
- Stolyarova A, Coop G, Przeworski M. 2025. The distribution of highly deleterious variants across human ancestry groups. *Proc Natl Acad Sci U S A.* 122:e2503857122. <https://doi.org/10.1073/pnas.2503857122>.
- The Gene Ontology Consortium et al. 2023. The gene ontology knowledgebase in 2023. *Genetics.* 224:iyad031. <https://doi.org/10.1093/genetics/iyad031>.
- The GTEx Consortium. 2020. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science.* 369:1318–1330. <https://doi.org/10.1126/science.aaz1776>.
- Vincent EE et al. 2015. Mitochondrial phosphoenolpyruvate carboxykinase regulates metabolic adaptation and enables glucose-independent tumor growth. *Mol Cell.* 60:195–207. <https://doi.org/10.1016/j.molcel.2015.08.013>.
- von Boehmer H. 1994. Positive selection of lymphocytes. *Cell.* 76:219–228. [https://doi.org/10.1016/0092-8674\(94\)90330-1](https://doi.org/10.1016/0092-8674(94)90330-1).
- Wall JD et al. 2023. South Asian medical cohorts reveal strong founder effects and high rates of homozygosity. *Nat Commun.* 14:3377. <https://doi.org/10.1038/s41467-023-38766-1>.
- Weghorn D et al. 2019. Applicability of the mutation–selection balance model to population genetics of heterozygous protein-truncating variants in humans. *Mol Biol Evol.* 36:1701–1710. <https://doi.org/10.1093/molbev/msz092>.
- White J. 2020. PubMed 2.0. *Med Ref Serv Q.* 39:382–387. <https://doi.org/10.1080/02763869.2020.1826228>.
- Wilson PD. 1997. Epithelial cell polarity and disease. *Am J Physiol Renal Physiol.* 272:F434–F442. <https://doi.org/10.1152/ajprenal.1997.272.4.F434>.
- Wright S, Dobzhansky T, Hovanitz W. 1942. Genetics of natural populations. VII. The allelism of lethals in the third chromosome of drosophila pseudoobscura. *Genetics.* 27:363. <https://doi.org/10.1093/genetics/27.4.363>.
- Xiao Q, Lauschke VM. 2021. The prevalence, genetic complexity and population-specific founder effects of human autosomal recessive disorders. *npj Genom Med.* 6:41. <https://doi.org/10.1038/s41525-021-00203-x>.
- Xue S et al. 2024. Dysregulation of phosphoenolpyruvate carboxykinase in cancers: a comprehensive analysis. *Cell Signal.* 120:111198. <https://doi.org/10.1016/j.cellsig.2024.111198>.
- Yengo L, Wray NR, Visscher PM. 2019. Extreme inbreeding in a European ancestry sample from the contemporary UK population. *Nat Commun.* 10:3719. <https://doi.org/10.1038/s41467-019-11724-6>.
- Young CL et al. 2023. A maternal germline mutator phenotype in a family affected by heritable colorectal cancer. <https://doi.org/10.1101/2023.12.08.23299304>.
- Zeng T, Spence JP, Mostafavi H, Pritchard JK. 2024. Bayesian estimation of gene constraint from an evolutionary model with gene features. *Nat Genet.* 56:1632–1643. <https://doi.org/10.1038/s41588-024-01820-9>.
- Zschocke J, Byers PH, Wilkie AO. 2023. Mendelian inheritance revisited: dominance and recessiveness in medical genetics. *Nat Rev Genet.* 24:442–463. <https://doi.org/10.1038/s41576-023-00574-0>.

Editor: J. Novembre