


Mutation–selection–drift balance models of complex diseases

Jeremy J. Berg ^{1,2,*} Xinyi Li,² Kellen Riall,² Laura K. Hayward,³ Guy Sella^{4,5,*}

¹Department of Human Genetics, The University of Chicago, Chicago, IL 60637, United States

²Committee on Genetics, Genomics and Systems Biology, The University of Chicago, Chicago, IL 60637, United States

³Institute of Science and Technology Austria Klosterneuburg 3400, Lower Austria, Austria

⁴Department of Biological Sciences, Columbia University, New York, NY 10027, United States

⁵Program for Mathematical Genomics, Columbia University, New York, NY 10027, United States

*Corresponding authors: Jeremy J. Berg, Department of Human Genetics, University of Chicago, 920 E 58th St CLSC, Chicago, IL 60637, USA. Email: jjberg@uchicago.edu; Guy Sella, Department of Biological Sciences, Columbia University, 1212 Amsterdam Ave, Fairchild Center, New York, NY 10027, USA. Email: gs2747@columbia.edu

Genetic variation that influences complex disease susceptibility is introduced into the population by mutation and removed by natural selection and genetic drift. This mutation–selection–drift balance (MSDB) shapes the prevalence of a disease and its genetic architecture. To date, however, MSDB has been modeled only for monogenic (Mendelian) diseases. Here, we develop an MSDB model for complex disease susceptibility: we assume that genotype relates to disease risk according to the canonical liability threshold model and that the selection on variants affecting risk stems from the fitness cost of the disease. We focus on diseases that are highly polygenic, entail a substantial fitness cost, and are neither extremely common in the population nor exceedingly rare. The comparison of model predictions with genome-wide association studies and other observations in humans indicates that common genetic variation affecting complex disease susceptibility is little affected by directional selection and instead shaped by pleiotropic stabilizing selection on other traits. In turn, directional selection may exert a more substantial effect on rare, large-effect variants. Our results also suggest that current estimates of disease heritability are likely biased. The model thus provides a better understanding of the evolutionary processes that shape the architecture and prevalence of complex diseases.

Keywords: mutation; selection; genetic drift; complex disease

Introduction

A central goal of population genetics is to understand how evolutionary processes shape the prevalence of genetic diseases and the population distribution of their underlying genetic variants. This question is of particular interest in humans. Since the late 20th century, we have learned a lot about the genetic basis of simple (Mendelian) diseases and the frequencies of their underlying variants in human populations (Jobling et al. 2013, Ch. 16). We also have long-standing models for the evolutionary processes that generate and maintain these diseases (Haldane 1927; Fuller et al. 2019), whose predictions are in qualitative, if not quantitative, agreement with empirical observations (see, e.g. Amorim et al. 2017).

Most common genetic diseases in humans (e.g. with a prevalence of $\geq 0.1\%$) are complex (Jobling et al. 2013, Ch. 17), however, and it is only over the past decade or so that genome-wide association studies (GWAS) have begun to reveal their genetic basis (The Wellcome Trust Case Control Consortium 2007; Trubetskoy et al. 2022). These studies have now identified many thousands of robust associations between genetic variants and many diseases, and in so doing, have begun to uncover the numbers, effect sizes, and frequencies of the variants underlying disease

risk—henceforth the “genetic architecture” of complex diseases (Abdellaoui et al. 2023). Yet we still lack a good understanding of the evolutionary processes that shape the architecture and prevalence of complex diseases.

The discoveries from GWAS shed some light on these processes. Notably, they reveal that variant effects on disease risk are negatively correlated with their minor allele frequency, indicating that natural selection acts to remove genetic variation affecting disease risk and that the strength of selection on variants increases with their effect on risk (Schoech et al. 2019; Zeng et al. 2021). Additionally, many of the significant associations in GWAS of diseases are common (see, e.g. Trubetskoy et al. 2022) indicating that for much of the variation affecting disease risk, the effects of selection on variant frequencies are comparable to those of random genetic drift. It further appears that variation in the risk of developing common complex diseases is thinly spread among many thousands of segregating variants that are widely distributed across the genome (Loh et al. 2015; Shi et al. 2016; Boyle et al. 2017). This extreme polygenicity, alongside evidence for selection and drift that lead to the removal of variation, implies that genetic variation is continually replenished by mutations at numerous sites across the genome. Thus,

Received on 30 May 2025; accepted on 16 September 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of The Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

complex disease prevalence and architecture are shaped by a balance between mutation, natural selection, and genetic drift.

Existing models for mutation–selection–drift balance (MSDB) come in 3 flavors. The first is the classic model for simple Mendelian diseases introduced by Danforth and Haldane and extended by Muller, Kimura, and others (Danforth 1923; Haldane 1927, 1937; Muller 1950; Crow 1958; Kimura, Maruyama, and Crow 1963; Clark 1998). In its most basic form, the model assumes that mutations arising at a single gene cause the disease in either heterozygotes (the dominant case) or homozygotes (the recessive case) and that the disease markedly reduces individual fitness (see, e.g. Gillespie 2004). The fitness cost of the disease induces strong selection against the alleles that cause it, leading to their loss from the population. The model describes the prevalence of the disease, the frequency of the underlying alleles, and the genetic load as a function of the mutation rate and fitness cost. This model and its generalizations, however, are not applicable to complex diseases, because the risk of developing complex diseases arises from the contribution of many variants (and effects of the environment), which, in turn, generates a less obvious relationship between the fitness cost of the disease and selection on its underlying variants.

The second flavor of MSDB models relates selection on complex traits to the selection acting on the many variants that affect these traits. They do not focus on disease risk, however, but on quantitative (continuous) traits, assuming that these traits are subject to stabilizing selection, i.e. that traits have an optimal value and individual fitness declines continuously with displacement from it (Robertson 1956; Lande 1975; Keightley and Hill 1988; Simons et al. 2018). Typically, these models further assume that mutations are equally likely to increase or decrease trait values and that the effect sizes in either direction have the same distribution, an assumption known as symmetric mutation (but see Waxman and Peck 2003; Zhang and Hill 2008; Charlesworth 2013a, 2013b). MSDB models of quantitative, complex traits have been invaluable in studying the processes that maintain heritable variation in complex traits. More recently, they have been used to study the genetic architecture of complex traits and to interpret the results of human GWAS (Simons et al. 2018; O'Connor et al. 2019; Zeng et al. 2021; Simons et al. 2025; Spence et al. 2024). It is not obvious that they apply to complex diseases, however.

Indeed, some diseases, for example hypertension or obesity, are defined in terms of underlying quantitative traits that exceed a threshold value, and these underlying traits may well be subject to stabilizing selection and have symmetric mutation. Other diseases, for instance type 2 diabetes, may reflect a discrete biological dysfunction, such as a breakdown of homeostasis (Alon 2023). In such cases, we might expect the effects of the disease on fitness to be discrete and selection to be directional, in always acting to reduce disease risk, and mutations may therefore tend to increase disease risk.

The third flavor of MSDB models includes all these elements, while considering fitness rather than disease risk as the focal trait. These models were introduced to study the fitness burden of deleterious mutations—the genetic load—in natural populations (Kimura and Maruyama 1966; King 1966; Kondrashov 1995) and were later related to evolutionary advantages of sex (Kondrashov 1982, 1988). They typically assume that fitness drops sharply around some threshold number of deleterious mutations or threshold “liability” (sometimes referred to as “fitness potential”; Milkman 1978; Kondrashov 2018), which arises from additive (weighted) contributions over all the deleterious alleles that an individual carries. With fitness always decreasing with an increasing number of deleterious alleles (or with increasing liability),

selection is always directed against these alleles. Under such directional selection, loci tend to be fixed for beneficial alleles, and therefore, mutations at these loci tend to be deleterious. With some modifications, these models could be framed as models of complex diseases that are similar to the model we present below.

In their existing form, however, such models cannot be related to the architecture and prevalence of complex disease. Indeed, some of the models (Kimura and Maruyama 1966; Kondrashov 1982, 1984) assume that the variants that underlie fitness are all subject to strong selection (i.e. selection that is much stronger than genetic drift), an assumption that contradicts what we have learned from GWAS of complex diseases (see above). Other studies rely on rough approximations to describe weakly selected variation (e.g. Kondrashov 1995) or assume a mapping between liability and fitness that is inconsistent with disease models (Charlesworth 1990, 2013b). These assumptions do not allow the genetic architecture and prevalence of complex diseases to be related to their underlying evolutionary parameters.

Motivated by these considerations, we develop an MSDB model of complex disease risk and solve it for the genetic architecture of the disease and its prevalence. Similar to classic models of simple (Mendelian) diseases and analogous to models of genetic load, we assume that directional selection always acts to reduce disease risk and that the disease state is discrete. Similar to models of complex quantitative traits and analogous to models of genetic load, we assume that disease risk arises from the joint effects of many variants and environmental effects. Unlike models of genetic load, we consider accurate approximations for the behavior of weakly selected variation affecting disease risk. By combining these features, we are able to describe the expected genetic architecture of complex diseases and their prevalence and to contrast our predictions with the findings of GWAS in humans.

The model

We use the canonical model for binary traits in human and quantitative genetics—the liability threshold model—to relate an individual’s genotype with their risk of developing a disease (Wright 1926, 1934; Lush et al. 1948; Dempster and Lerner 1950; Falconer 1965). Liability is a quantitative (continuous) trait that cannot be observed directly. When an individual’s liability, Z , exceeds a threshold, T , they will develop or have the disease (Fig. 1a).

An individual’s liability relates to their genotype through the standard additive model of quantitative complex traits (Falconer and Mackay 1995). Namely, we assume that the number of genomic sites affecting liability (i.e. the target size) is very large, $L \gg 1$, and that an individual’s liability is given by

$$Z = G + E = \sum_{\ell=1}^L a_{\ell} g_{\ell} + E, \quad (1)$$

where G is the genetic contribution, which is the sum of contributions over sites; a_{ℓ} is the effect size of the liability increasing allele at site ℓ ; g_{ℓ} is the number of copies of that allele at that site; and $E \sim N(0, V_E)$ is the environmental contribution.

This model implies that an individual’s total liability is normally distributed around their genetic liability, with the variance arising from the environmental contribution, namely that $Z|G \sim N(G, V_E)$. An individual’s probability of developing the disease, their genetic risk, is therefore

$$R|G = 1 - \Phi(T|G, V_E), \quad (2)$$

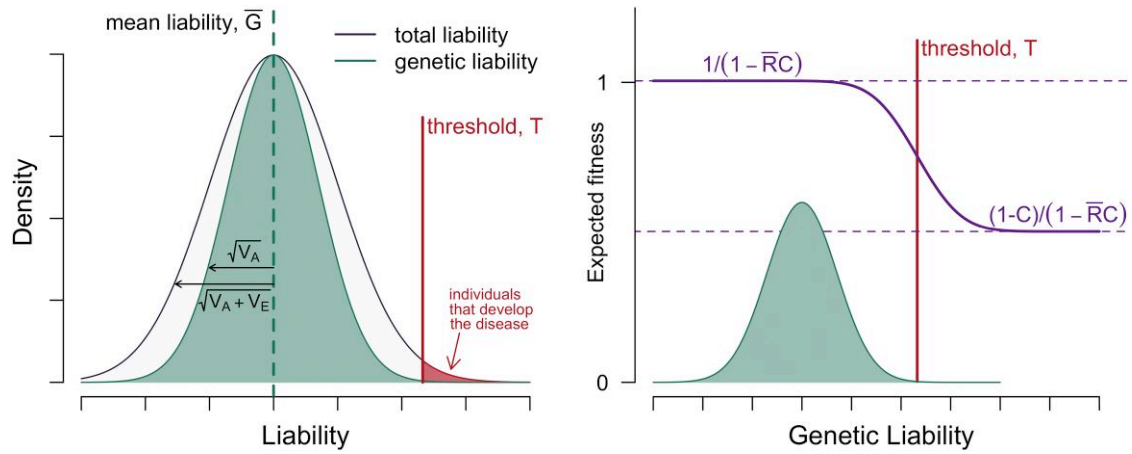


Fig. 1. The model. a) The liability distribution in the population. Individuals whose liability exceeded the threshold develop the disease (in red). b) The fitness landscape. In these illustrations, we assume that the genetic ability distribution is normal, that the heritable variance in liability is $h^2 = 1/2$, the fitness cost $C = 1/2$, and the prevalence $\bar{R} = 0.01$.

where $\Phi(\cdot|G, V_E)$ is the normal cumulative distribution function with mean G and variance V_E .

We model fitness by assuming that the disease entails a fitness cost C , such that an individual without the disease has fitness 1 and one with the disease has fitness $1 - C$. The mean fitness of a population with disease prevalence \bar{R} is therefore $\bar{W} = 1 - C \cdot \bar{R}$, and the relative fitness associated with genetic liability G is

$$W|G = \frac{1 - C \cdot R|G}{1 - C \cdot \bar{R}} \quad (3)$$

(with $R|G$ given in Equation 2).

Figure 1b shows the resulting fitness landscape. Fitness plateaus at $1/(1 - C \cdot \bar{R})$ at low genetic liability and at $(1 - C)/(1 - C \cdot \bar{R})$ at high genetic liability, and the width of the transition between plateaus reflects the variance of the environmental contribution. This drop-off in fitness leads to selection to reduce the mean genetic liability of the population.

For simplicity, we assume that the genomic sites affecting liability are biallelic, and without loss of generality, we set the liability scale such that at a given site ℓ , the low-liability allele contributes 0 and the high-liability allele contributes liability $a_\ell > 0$. We denote the distribution of effects across sites by $g(a)$ and its mean by \bar{a} . In these terms, the possible values of genetic liability range between 0 and $2L\bar{a}$ (when all sites are fixed for the low- or high-liability alleles, respectively). We assume that the threshold is in the lower half of the liability range, i.e. $T < L\bar{a}$ (otherwise, the disease would rarely or never manifest and would have little or no evolutionary effect).

Each generation, mutation introduces new variants that affect liability into the population. We assume that the 2 possible alleles mutate to one another with probability u per gamete per generation. We further assume, for simplicity, that the rate of mutational input per site in the population is sufficiently low for us to rely on the infinite-sites approximation (specifically, we assume that $\theta/2 = 2Nu \ll 1$, where N is the population size). In this approximation, segregating derived alleles are assumed to have arisen from a single mutation, and the number of mutations per gamete per generation follows a Poisson distribution with mean Lu . The infinite-sites approximation is standard and sensible in many contexts in humans (Harpak et al. 2016; Schraiber et al. 2024).

Table 1. Notation used in the main text.

Symbol	Meaning
Z	Liability
G	Genetic liability
E	Environmental liability
T	Threshold
R	Genetic risk
L	Mutational target size
u	Per-site mutation rate
N	Population size
$\theta/2 = 2Nu$	Per-site population scaled mutation rate
C	Fitness cost of disease
W	Fitness
V_A	Genetic variance
V_E	Environmental variance
V_P	Total phenotypic variance
h^2	Heritability of liability
g	Genotype at a site
a	Liability effect of an allele
$g(a)$	Distribution of liability effect sizes
$\partial_R(a)$	Risk effect of alleles with liability effect a
$s(a)$	Selection coefficient of alleles with liability effect a
x	Frequency of risk allele at a site
$F(Z)$	Probability that an individual's liability exceeds Z
$f(Z)$	Probability density of liability at Z
$p_+(a)/p_-(a)$	Proportion of sites with effect a fixed for risk-increasing/risk-decreasing allele and risk decreasing sites respectively
$\gamma(a)$	Scaled selection coefficient of a site with liability effect a
$b(a)$	Fixation bias toward risk-decreasing alleles
$v(a)$	Contribution to liability variance per unit diploid mutation rate
$\Delta_U \bar{G}$	Mutational increase in mean liability
$\Delta_S \bar{G}$	Selection response of mean liability
\bar{b}	Mean fixation bias weighted by effect size
b_T	Threshold bias
ϕ_T	Standardized threshold density
p_s/p_l	Fraction of sites in 2 effect models that have small/large effects

The population dynamics follow the standard model of a diploid, panmictic population of constant size N , with nonoverlapping generations. In each generation, parents are randomly chosen to reproduce with probabilities proportional to their fitness (Equation 3), i.e. Wright-Fisher sampling with fertility selection, followed by mutation, free recombination (i.e. no linkage), and Mendelian segregation. Our notation is summarized in Table 1.

Scope

We focus on diseases that are highly polygenic, have a substantial fitness cost, and are not extremely common or exceedingly rare. Specifically, our analysis should apply to diseases with a target size $L \gtrsim 10^5$ (noting that the target size is typically much greater than the polygenicity), fitness cost $C \gtrsim 0.1$, and prevalence $10\% \gtrsim \bar{R} \gtrsim 0.1\%$. These assumptions plausibly encompass most common complex diseases in humans. Moreover, within these parameter ranges, the liability threshold model on which we rely and other standard models of complex diseases in human statistical genetics are practically interchangeable (Slatkin 2008; Wray and Goddard 2010); these include Risch's (1990) multiplicative model used as premise in linkage studies and the logistic model used in case-control GWAS (Sham and Purcell 2014).

We focus on the model's behavior at MSDB (i.e. at equilibrium). For the type of diseases we consider, most of the liability distribution falls below the threshold, with only a small tail above it (Fig. 1a). For simplicity, we assume that the liability distribution is well approximated by its stationary distribution and ignore stochastic fluctuations around this stationary distribution. Under our assumptions (notably of high polygenicity), this is a sensible assumption that should not have a substantial effect on our results.

Simulations

We validate our analytic results using simulations. The simulations are implemented in SLiM (version 3.6; Haller and Messer 2019) and realize the models with 1 or 2 liability effect sizes. We initialize simulations with each site fixed for one of its alleles. The proportions of sites fixed for each allele are set according to our analytic expectations at MSDB, which we derive below (otherwise, reaching MSDB would take prohibitively long); the initial population thus consists of genetically identical, homozygous individuals at all sites. We check that our initial fixed state approximates the fixed state at MSDB by recording fixations in 25 replicate simulations over $500N$ generations and comparing both the net and total fixation fluxes with the analytic expectations (Supplementary Section 2.5; Supplementary Fig. 14). We use a separate set of simulations to measure quantities of interest at MSDB. We run each simulation for a burn-in period of $10N$ generations to allow genetic variation to approach the steady state at MSDB. We then run each simulation for an additional $25N$ generations and sample the population every $0.05N$ generations, amounting to 500 samples per simulation. In most cases, we run 6 replicate simulations for each parameter setting and estimate quantities of interest by averaging over $500 \times 6 = 3000$ samples. For more details about the simulations, see Supplementary Section 6.

Results

The population dynamics and genetic architecture at individual sites

The dynamics at a site

The dynamics at a site can be described in terms of the first 2 moments of change in allele frequency in a single generation (Ewens 2004, Ch. 4). We calculate the moments by averaging the fitness of the 3 genotypes over genetic backgrounds and plugging these averages into the standard equation for the change in allele frequency at a single locus (Supplementary Section 1.1). We find that the expected change in frequency of an allele at frequency x that increases liability by a is well approximated by

$$E(\Delta x) = -s(a, x) \cdot x(1 - x). \quad (4)$$

The selection coefficient $s(a, x)$ takes an intuitive form: it equals the fitness cost of the disease, C , multiplied by the allele's effect on disease risk, $\delta_R(a, x)$, namely

$$s(a, x) \approx C \cdot \delta_R(a, x). \quad (5)$$

$\delta_R(a, x)$ is defined as the expected increase in individual disease risk caused by substituting a random liability-decreasing allele at this site by a liability-increasing one. As an aside, we note that the allele's effect on risk equals its (absolute) "population attributable risk" in epidemiology and can be translated into its odds ratio estimated in case-control GWAS (see Supplementary Section 1.5). The second moment of change in allele frequency is well approximated by the standard drift term

$$V(\Delta x) \approx x(1 - x)/2N. \quad (6)$$

To complete the description of these dynamics, we require the functional form of the risk effect $\delta_R(a, x)$. In Supplementary Section 1, we show that the risk effect is well approximated by the area under the liability distribution that is pushed over the liability threshold when the distribution is shifted by a on the liability scale (Fig. 2). We denote the probability density at liability Z by $f(Z)$ and the probability that liability exceeds liability Z by $F(Z)$. In these terms,

$$\delta_R(a, x) \approx F(T - a) - F(T) \quad (7)$$

We can therefore assume that the risk effect and selection coefficient do not vary with allele frequency and adjust our notation to $\delta_R(a)$ and $s(a)$, dropping the dependence on x .

The dependence of an allele's risk effect on its liability effect can be divided into 2 cases (Fig. 2). When the allele's effect on liability (a) is sufficiently small such that the corresponding shift to the liability distribution has a negligible effect on the

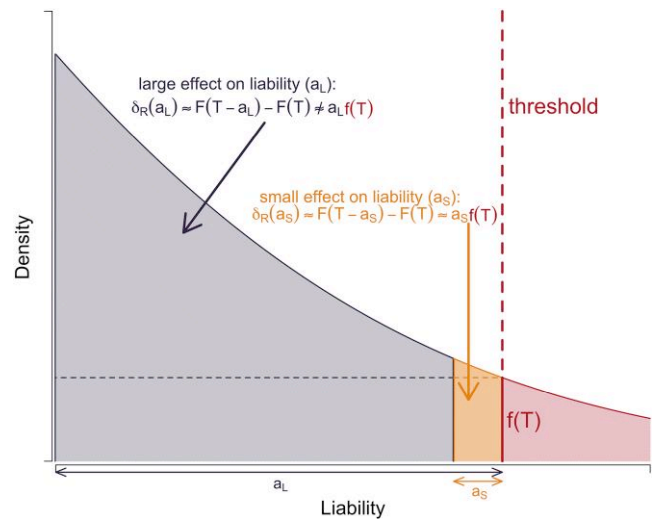


Fig. 2. The mapping between liability and risk effects. For small-effect sites, illustrated in yellow, the risk effect is approximately equal to the product of the liability effect and the threshold density. For large-effect sites, this linear approximation is inaccurate, and the risk effect must be approximated in terms of the difference in the areas in the tails of the liability distribution.

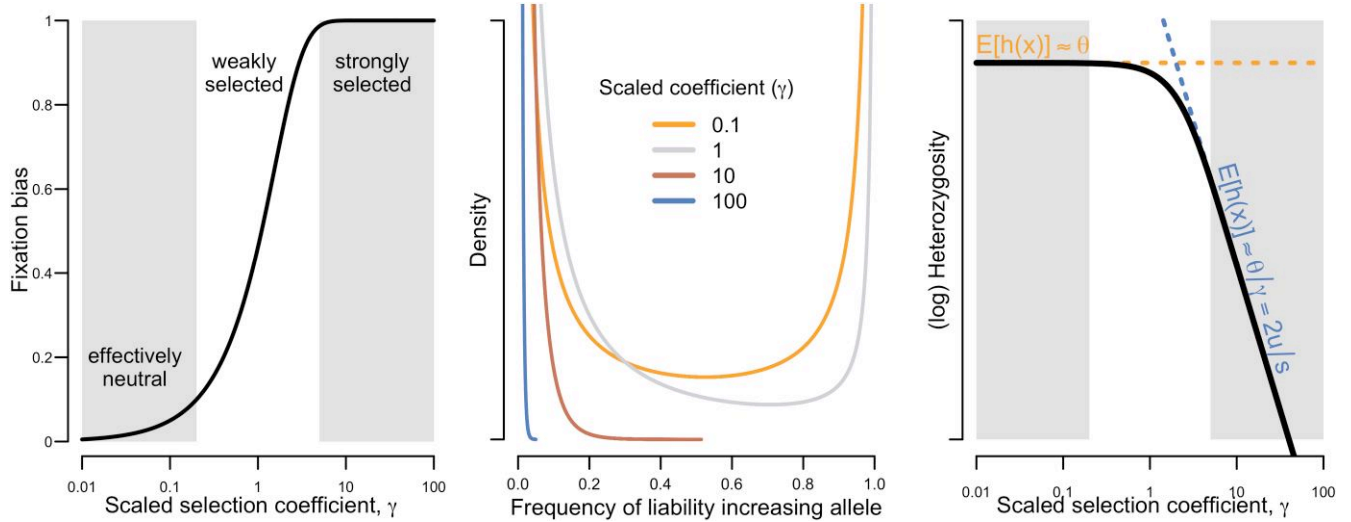


Fig. 3. The genetic architecture at MSDB. a) The fixation bias as a function of the scaled selection coefficient at a site. b) The site frequency spectrum. c) The expected genetic diversity as a function of the scaled selection coefficient.

probability density near the threshold (T), we can approximate the allele's effect on risk by the area of the rectangle with width a and height $f(T)$, i.e.

$$\delta_R(a, x) \approx F(T - a) - F(T) \approx a \cdot f(T) \quad (8)$$

(Kimura and Crow 1978; Milkman 1978). An allele is small in this sense if its effect is substantially smaller than the width of the phenotypic distribution, i.e. $a \ll \sqrt{V_P}$ (see Supplementary Section 1.3 and Supplementary Figs. 8 and 9).

In turn, an allele is large if its effect on liability is comparable to or greater than the width of the phenotypic distribution, i.e. $a \gtrsim \sqrt{V_P}$. When the effect of alleles on liability increases from small to large, the dependence of their risk effect δ_R on their liability effect (a) becomes superlinear (Fig. 2; but see Supplementary Fig. 1). In Supplementary Section 1.4, we show that for the kinds of diseases that we consider—that have a substantial fitness cost and are not exceedingly rare—alleles become strongly selected (i.e. $2Ns(a) \approx 2NC\delta_R(a) \gg 1$) before the superlinear dependence kicks in. This finding implies that we can divide alleles into 3 categories: small (in the sense of Equation 8) and weakly selected (i.e. $2Ns(a) \approx 2NC\delta_R(a) \lesssim 1$), small and strongly selected, and large and strongly selected.

Next, we consider the genetic architecture at MSDB. To this end, we only care about alleles' selection coefficients (rather than their effects on liability). The insensitivity of alleles' risk effect to their frequency allows us to treat their selection coefficients as constant, which, in turn, allows us to apply standard approximations to solve for quantities of interest throughout the range of allele effect sizes.

The fixed state

At MSDB, fixations are at detailed balance: for a given effect size, the rates of fixation of liability-increasing and liability-decreasing alleles at sites are equal, i.e. they balance each other out (Iwasa 1988; Sella and Hirsh 2005). The proportions of sites with effect a fixed for the risk-increasing and risk-decreasing alleles, $p_+(a)$ and $p_-(a) = 1 - p_+(a)$, respectively, therefore satisfy

$$p_+(a) \cdot \pi(s(a), \gamma(a)) = p_-(a) \cdot \pi(-s(a), -\gamma(a)), \quad (9)$$

where $\pi(s, \gamma) \approx 2s/(1 - e^{-2\gamma})$ is the fixation probability of a mutation with selection coefficient s and scaled selection coefficient $\gamma = 2Ns$ that arises at frequency $1/2N$ (Crow and Kimura 1970). We solve for the fixed state and represent the solution in terms of the bias toward risk-decreasing alleles

$$b(a) \equiv p_-(a) - p_+(a) = \frac{e^{2\gamma(a)} - 1}{e^{2\gamma(a)} + 1}. \quad (10)$$

In these terms, $p_+(a) = (1 - b(a))/2$ and $p_-(a) = (1 + b(a))/2$. The expected bias satisfies $1 > b(a) > 0$, because selection always favors the risk-decreasing allele.

The fixed state exhibits the 3 standard selection regimes (Fig. 3a). When selection is extremely weak and drift dominates, sites are equally likely to be fixed for the risk-increasing and risk-decreasing alleles, i.e. $b(a) \approx 0$. This effectively neutral regime occurs when $\gamma(a) \approx 2NC\delta_R(a) \ll 1$. At the other extreme, when selection is strong and dominates over drift, sites are always fixed for the risk-decreasing allele, i.e. $b(a) \approx 1$. This strong-selection regime occurs when $\gamma(a) \approx 2NC\delta_R(a) \gg 1$. In the weak-selection regime, when $\gamma(a) \approx 2NC\delta_R(a) \sim 1$, the fixed state transitions between the effectively neutral and strongly selected extremes, with the bias $b(a)$ increasing between 0 and 1 as selection becomes stronger.

With the fixed state biased toward risk-decreasing alleles (at all but effectively neutral sites), mutation is biased toward risk-increasing alleles. Like in the classic model of simple (Mendelian) diseases, this mutational asymmetry arises from the dynamics of the model rather than from assumptions about mutation (namely we assumed symmetric mutation between risk-increasing and risk-decreasing alleles).

Segregating sites

Figure 3b shows the frequency distribution of segregating, disease-increasing alleles at MSDB for several values of the population-scaled selection coefficient (see Supplementary Section 2.1 for derivation). With the fixed state biased toward alleles that decrease risk, derived, segregating alleles tend to

increase risk. Even neutral derived alleles, let alone derived alleles that increase risk, segregate at lower frequencies than ancestral ones. These effects explain why risk-increasing alleles tend to segregate below frequency $\frac{1}{2}$, and why this bias is stronger when selection is stronger. The predicted asymmetry between the frequencies of alleles that increase and decrease risk can be tested using data from GWAS (see Discussion and Koch et al. 2024).

Next, we consider diversity levels. We can calculate the expected diversity levels using the diffusion approximation (Crow and Kimura 1970; Ewens 2004). A variant with allele frequency x contributes $h(x) = 2x(1-x)$ to heterozygosity. To calculate the expectation per site, we multiply the rate at which mutations arise by the expected total contribution of an individual mutation during its sojourn in the population. Namely,

$$E(h(x)|\gamma) \approx 2Nu \cdot \int_0^1 h(\bar{x}) \cdot \tau(\bar{x}|\gamma) d\bar{x}, \quad (11)$$

with the sojourn time

$$\tau(x|\gamma) = \frac{2}{1 + e^{-2\gamma}} \frac{e^{-2\gamma x}}{x(1-x)} \quad (12)$$

defined such that $\tau(x|\gamma)dx$ is the expected number of generations that an allele with scaled selection coefficient γ spends between frequencies x and $x + dx$. In this way, we find the expected heterozygosity per site:

$$E(h(x)|\gamma(a)) \approx \theta \cdot b(a)/\gamma(a), \quad (13)$$

where $\theta = 4Nu$ and $b(a)$ is the fixed bias.

Diversity levels also exhibit the 3 standard selection regimes (Fig. 3c). In the effectively neutral regime (i.e. when $\gamma \ll 1$), heterozygosity is well approximated by the neutral expectation θ . In the strong selection regime (i.e. $\gamma \gg 1$), sites are always fixed for the risk-decreasing allele, implying that $b(a) \approx 1$ and that the derived, risk-increasing allele segregates at a low frequency, i.e. $x \ll 1$, so

$$E(h(x)|\gamma \gg 1) \approx E(2x|\gamma \gg 1) \approx \theta/\gamma = 2u/s, \quad (14)$$

which aligns with the classic expectation under mutation–selection balance. In the weak-selection regime (i.e. when $\gamma \sim 1$), diversity levels transition between these 2 extremes.

Contribution to variance

The last facet of architecture that we consider here is the contribution to additive variance in liability. Estimates of this contribution are used to assess how much of the heritable variance in disease risk arises from variants of small and large effects (see Discussion). The total genetic variance in liability will become important when we consider disease prevalence.

We can calculate the expected contribution to variance based on the expected heterozygosity (Equation 13). A variant with allele frequency x and liability effect a contributes $v(a, x) = 2a^2x(1-x) = a^2 \cdot h(x)$ to additive variance in liability. Therefore, the expected contribution per site is

$$E(v(a, x)) = a^2 \cdot E(h(x)|\gamma(a)) \approx a^2 \cdot \theta \cdot b(a)/\gamma(a) = 2u \cdot v(a), \quad (15)$$

where

$$v(a) = \frac{b(a)}{C} \frac{a^2}{\delta_R(\bar{a})} \quad (16)$$

is the contribution per unit diploid mutation rate.

We can calculate the total additive variance in liability by summing over sites and integrating over the distribution of effect sizes. Namely,

$$V_A = 2Lu \int_a v(a) \cdot g(a) da, \quad (17)$$

where L is the number of sites and g is the distribution of effect sizes. If we assume that all effect sizes are small, then $\delta_R(a) \approx a \cdot f(T)$ (Equation 8) and $v(a) \approx b(a) \cdot a/(C \cdot f(T))$. In this case, the total variance is well approximated by

$$V_A \approx (2Lu/Cf(T))\bar{a} \int_a b(a)(a/\bar{a})g(a)da = 2Lu\bar{a}\bar{b}/(Cf(T)), \quad (18)$$

where \bar{a} is the mean effect size and $\bar{b} \equiv \int_a b(a)(a/\bar{a})g(a)da$ is the mean fixation bias, with sites weighted by their effect sizes.

In Supplementary Section 2.3, we investigate how the contributions to liability- and risk-scale variance vary with variant effect sizes. We show that in our model, these variances do not exhibit the same kind of asymptotic “flattening” found in models of stabilizing selection (see Discussion and Supplementary Figs. 10 and 11, as well as Simons et al. 2018; O’Connor et al. 2019). We also describe how some simple summaries of asymmetry in the architecture at individual sites can be calculated (Supplementary Section 2.4; Supplementary Figs. 12 and 13). Like the summaries considered here, all these summaries can be described in terms of a , θ , and $\gamma(a)$.

The mapping between liability effects and selection coefficients

A phenotypic perspective on MSDB

At MSDB, mutation is biased toward risk-increasing alleles, causing a mutational increase in mean genetic liability each generation; we denote it by $\Delta_U \bar{G}$. The mutational bias is balanced by an equal but opposing selection response that we denote by $\Delta_S \bar{G}$. Thus, at MSDB, $\Delta_U \bar{G} + \Delta_S \bar{G} = 0$.

Here, we focus on the mutational bias. We first approximate the mutational bias based on the fixed state. In this approximation, all risk-increasing mutations occur at sites fixed for the risk-decreasing allele and vice versa. As above, we denote the proportions of sites with effect a fixed for the risk-increasing and risk-decreasing alleles at MSDB by $p_+(a)$ and $p_-(a) = 1 - p_+(a)$, respectively. We then approximate the mutational bias as

$$\Delta_U \bar{G} \approx 2Lu \int_a (p_-(a) - p_+(a)) a g(a) da, \quad (19)$$

where $g(a)$ is the distribution of liability effects across sites. Recalling that the fixation bias $b(a) \equiv p_-(a) - p_+(a)$, we can express the mutational bias as

$$\Delta_U \bar{G} \approx 2Lu \int_a b(a) a g(a) da = 2Lu\bar{a} \int_a b(a)(a/\bar{a})g(a)da = 2Lu\bar{a}\bar{b}, \quad (20)$$

where \bar{a} is the mean effect size and \bar{b} is the mean fixation bias, weighted by effect sizes.

We can also express the mutational bias in terms of the mean genetic liability \bar{G} . In the fixed state approximation, the mean liability is

$$\bar{G} \approx 2L \int_a p_+(a) a g(a) da = 2L \int_a (1 - p_-(a)) a g(a) da, \quad (21)$$

where we have set the range of genetic liability scale between 0 and $2L\bar{a}$, with low-liability alleles contributing 0. Substituting Equations 21 into Equation 19, we find that

$$\Delta_U \bar{G} = 2Lu\bar{a} \left(1 - \frac{\bar{G}}{L\bar{a}}\right), \quad (22)$$

where the comparison with Equation 20 implies that $\bar{b} \approx \left(1 - \frac{\bar{G}}{L\bar{a}}\right)$. In Supplementary Section 3, we show that Equation 22 is exact when we relax the fixed-state approximation and account for segregating genetic variation.

Under our modeling assumptions, the distance between the mean genetic liability (\bar{G}) and the liability threshold (T) is tiny relative to the possible range of genetic liability ($2L\bar{a}$), i.e. $(T - \bar{G}) \ll 2L\bar{a}$. To understand why, we consider the case without an environmental contribution to liability. (An environmental contribution reduces only the distance between the mean genetic liability and the threshold, because it reduces the efficacy of selection on individual sites.) Without an environmental contribution, our assumption of a low mutation rate per site ($\theta/2 = 2Nu \ll 1$) implies that the scale of variation in liability among individuals is much smaller than the scale of possible genetic liabilities (i.e. $\sqrt{V_G} \ll 2L\bar{a}$). In turn, our assumption that the disease prevalence is not exceedingly small requires the distance between the mean genetic liability and the threshold to be on the scale of the genetic variation (i.e. $(T - \bar{G})/\sqrt{V_G} \sim 1$). It therefore follows that $(T - \bar{G})/2L\bar{a} \ll 1$.

This condition allows us to approximate the mutational bias in terms of the position of the threshold. Specifically, given that $\bar{G}/L\bar{a} \approx T/L\bar{a}$, we find that

$$\Delta_U \bar{G} = 2Lu\bar{a} \left(1 - \frac{\bar{G}}{L\bar{a}}\right) \approx 2Lu\bar{a} \left(1 - \frac{T}{L\bar{a}}\right) = 2Lu\bar{a} \cdot b_T, \quad (23)$$

where $b_T \equiv 1 - \frac{T}{L\bar{a}}$ measures the position of the threshold relative to the middle of the liability scale; we henceforth refer to it as the threshold bias. Equation 23 shows that the threshold bias determines the mutational bias. Moreover, comparing Equations 20 and 23, we find that

$$\bar{b} = \int_a b(a) (a/\bar{a}) g(a) da \approx b_T, \quad (24)$$

where our assumption that $T < L\bar{a}$ (see Model section) is equivalent to assuming that $b_T > 0$ and ensures that $\bar{b} > 0$. Thus, the threshold bias also determines the mean fixed bias at MSDB, which reflects the strength of selection acting on individual sites (Equation 10).

Selection at sites with small effects

Equation 24 also tells us how the strength of selection on sites relates to their effects on liability, so long as these effects are small. To understand how, we first consider a simple case in which all sites have the same liability effect size a and therefore the same scaled selection coefficient $\gamma(a)$. When we express the fixation bias in terms of the scaled selection coefficient (Equation 10), Equation 24 becomes

$$\bar{b} = b(a) = \frac{e^{2\gamma(a)} - 1}{e^{2\gamma(a)} + 1} \approx b_T. \quad (25)$$

Solving for the scaled selection coefficient, we find that

$$\gamma(a) \approx \frac{1}{2} \ln \frac{1 + b_T}{1 - b_T}. \quad (26)$$

Thus, the position of liability threshold b_T determines the scaled selection coefficient at MSDB $\gamma(a)$ (Fig. 4a). Equation 26 applies so long as the threshold is not extremely close to 0 (specifically, $T/(L\bar{a}) = 1 - b_T \gg 1/L$), such that some sites are fixed for the liability-increasing allele. For this to be the case, selection cannot be too strong, implying that our small-effect-size approximation (Equation 8) applies and that $\gamma(a) \approx 2NC\delta_R(a) \approx (2NCa) \cdot f(T)$. Consequently, the threshold density at MSDB is

$$f(T) \approx \frac{1}{4NCa} \ln \frac{1 + b_T}{1 - b_T}. \quad (27)$$

Thus, given the compound evolutionary parameter $4NCa$, the required strength of selection is attained by adjusting the population's liability distribution, such that the threshold density $f(T)$ satisfies Equation 27.

Next, we consider the general case with a distribution of effect sizes. In this case, when we express Equation 24 in terms of scaled selection coefficients, we find that

$$\bar{b} = \int_a b(a) (a/\bar{a}) g(a) da = \int_a \frac{e^{2\gamma(a)} - 1}{e^{2\gamma(a)} + 1} (a/\bar{a}) g(a) da \approx b_T. \quad (28)$$

If we assume the small-effect approximation for $\gamma(a)$, the equation becomes

$$\bar{b} = \int_a \frac{\text{Exp}(4NCf(T) \cdot a) - 1}{\text{Exp}(4NCf(T) \cdot a) + 1} \frac{a}{\bar{a}} g(a) da \approx b_T. \quad (29)$$

We can solve this equation numerically using a line search for the threshold density $f(T)$ given the distribution g , $4NC$, and b_T . Importantly, given the $f(T)$ that solves Equation 29, $\gamma(a) \approx (2NCa) \cdot f(T)$ solves Equation 28.

The solution for the threshold density divides sites into 3 kinds. On the high end of effect sizes, sites are strongly selected, with $(2NCa) \cdot f(T) \gg 1$, and thus fixed for the liability-decreasing alleles; these sites all contribute $b \approx 1$ to attaining $\bar{b} \approx b_T$. On the lower end of effect sizes, sites are effectively neutral, with $(2NCa) \cdot f(T) \ll 1$, and thus equally likely to be fixed for the liability-increasing and liability-decreasing alleles; these sites all contribute $b \approx 0$ to attaining $\bar{b} \approx b_T$. In between these ends, sites are weakly selected, with $(2NCa) \cdot f(T) \sim 1$, and their fixed state is highly sensitive to the threshold density, with the fixed bias b ranging between 0 and 1.

If the distribution of liability effects g is highly concentrated around \bar{a} , the solution resembles the case with a single effect size. MSDB is attained by adjusting the threshold density such that most sites—with effects near \bar{a} —are weakly selected, and $b(\bar{a}) \approx b_T$. In the other extreme, with liability effects thinly distributed over several orders of magnitude, only a small proportion of sites would fall in the intermediate, weakly selected range of effect sizes. In this case, the threshold density at MSDB divides the range of effect sizes such that the fixed bias from strongly selected sites matches the threshold bias, i.e. such that $\bar{b} \approx \int_{(2NCf(T) \cdot a) \gg 1} 1 \cdot (a/\bar{a}) g(a) da \approx b_T$. In Supplementary Section 4, we investigate how variation in the distribution of liability effects, g , and relative position of the liability threshold, b_T , would affect

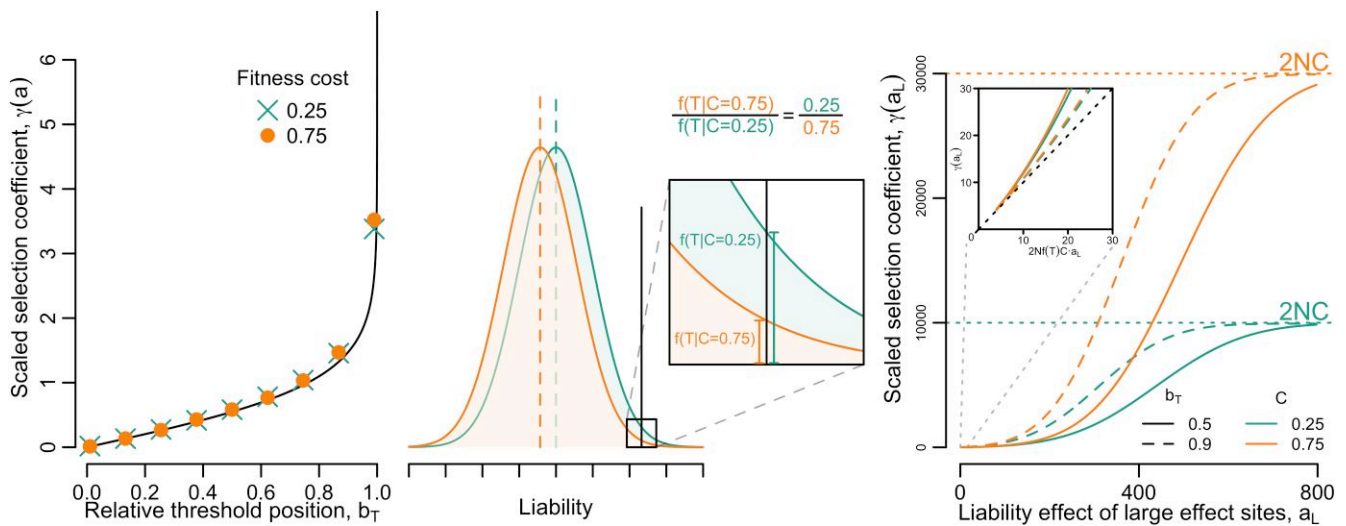


Fig. 4. The mapping between liability and fitness effects. a) The threshold position determines the scaled selection coefficients in the model with a single effect size. The solid line depicts the analytic approximation (Equation 6) and the circles and crosses depict averages in simulations with specified fitness costs (see Model and Supplementary Section 6 for details). Scaled selection coefficients in simulations were estimated by computing the average of the risk effect, $\delta_R(a)$, over many sites and generations and multiplying by $2NC$. b) Selection on small-effect sites is insensitive to the cost of the disease. When the fitness cost, C , increases, mean liability is pushed down to reduce the threshold density, $f(T)$, such that $2NCf(T)$ and thus scaled selection coefficients at small-effect sites remain invariant. c) The mapping between liability and scaled selection effects at large-effect sites. The results shown are based on numerically solving the model with 2 effect sizes (see Supplementary Section 6 for details), setting $N = 20,000$, $L = 1.5 \times 10^7$, and $u = 10^{-8}$ per site per generation, with fractions $p_s = 0.9995$ of small-effect sites and $p_l = 0.0005$ of large-effect sites, varying a_l with $a_s = 1$, and using the specified values of b_T and C .

the solution of Equation 29, assuming most sites have small effects (Supplementary Figs. 15 and 16).

Here, we focus on what the solution of Equation 29 tells us about the mapping between liability effects and selection coefficients. As we already know, the approximation $\gamma(a) \approx (2NCa) \cdot f(T)$ breaks down when sites are sufficiently large (see, e.g. Figure 2). This does not affect the solution to Equation 29 because selection at these sites is sufficiently strong (i.e. $\gamma(a) \gg 1$) to maximize their fixed bias (i.e. $b(a) \approx 1$) regardless of the exact form relating their liability effect sizes and selection coefficients. This reasoning clarifies that Equation 29 is insensitive to, and uninformative about, the strength of selection acting on sites with sufficiently large effects.

In contrast, the strength of selection acting on sites with small effects follows from Equation 29. When effect sizes are sufficiently small, the strength of selection is well approximated by $\gamma(a) \approx (2NCa) \cdot f(T)$. As we already noted, for diseases that have a substantial fitness cost and are not exceedingly rare, sites with such small-effect liabilities range from being effectively neutral to being strongly selected. Selection on sites in this range is determined by the relative position of the liability threshold, b_T , and the distribution of effects, g , where the mapping between effect sizes and scaled selection coefficients is attained by adjusting the threshold density, $f(T)$.

The genetic architecture at sites with small effects

Our results imply that scaled selection coefficients at sites with small effects are insensitive to the fitness cost, C ; environmental variance, V_E ; and population size, N . Figure 4b illustrates the effects of an increase in fitness cost. In this case, the mean genetic liability (\bar{G}) is pushed farther below the threshold to reduce the threshold density ($f(T)$), leaving $4NCf(T)$ unchanged and maintaining the same mapping between liability effect sizes and scaled selection coefficients. An increase in population size or decrease in environmental variance acts similarly, although they

also affect the total variance in liability. Because the disease is highly polygenic, changes in mean genetic liability are achieved by tiny changes to the fixed state across sites, with negligibly small effects on the scaled selection coefficients.

As we described earlier, the architecture at sites depends only on a , θ , and $\gamma(a)$. For sites with small effects, the dependence on γ translates into a dependence on the threshold bias, b_T , and the distribution of liability effects, g . In turn, the architecture at sites with small effects is insensitive to the fitness cost of the disease and the environmental variance, while the population size affects only the total number of segregating sites, but not the distribution of their allele frequencies. It follows that dynamic changes to the fitness cost or environmental variance would also have a negligible effect on the architecture, but it would take on the order of N_e generations for the difference in the number of segregating sites to propagate to all allele frequencies after a change in population size.

Selection on and genetic architecture of sites with large effects

Large effect sites span a wide range of liability effect sizes, and the factors that determine the strength of selection acting on them vary within this range. At the lower end of this range, liability effect sizes are just above those of strongly selected, small-effect sites. Near this boundary, scaled selection coefficients are still strongly affected by the threshold density $f(T)$, as well as by derivatives of the liability distribution at the threshold (see, e.g. Figure 2). We would therefore expect selection on such large-effect sites to resemble selection on strongly selected, small-effect sites in being affected by the relative position of the threshold, b_T , and distribution of selection effects, g .

At the other end, we have sites with effect sizes that are large enough for a single copy of the risk-increasing allele to almost always cause the disease (e.g. when $a \gg \sqrt{V_P}$); these alleles are dominant, with full or nearly full penetrance. At this end, $\delta_R(a) \approx 1$, $s \approx C$, $\gamma \approx 2NC$, and the expected frequency of the risk-increasing

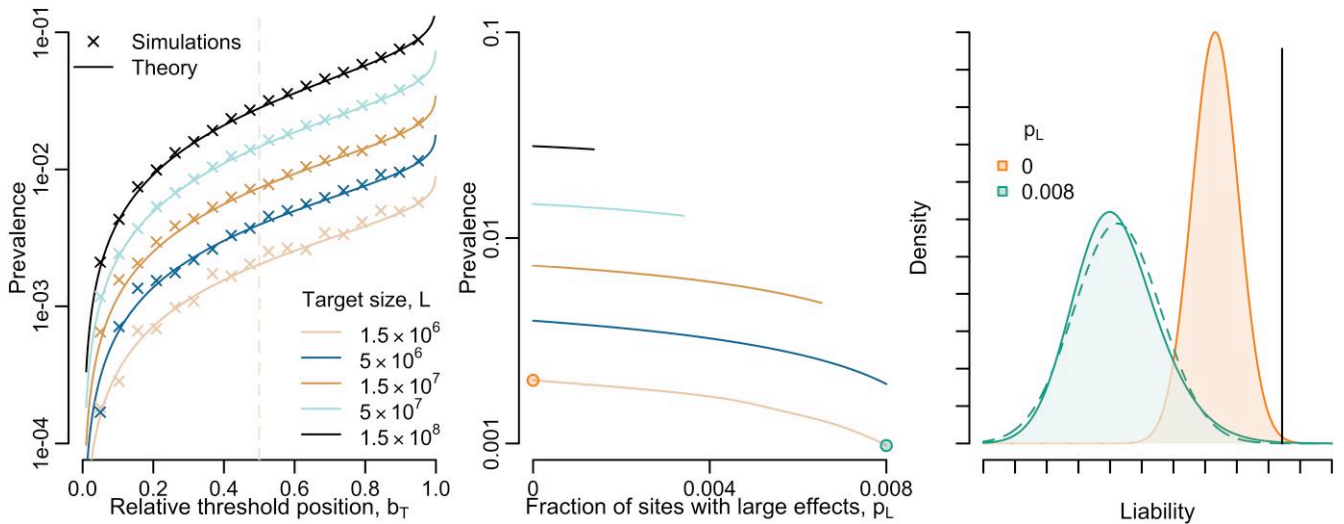


Fig. 5. Disease prevalence at MSDB. a) Prevalence vs threshold bias in the model with a single effect size. Analytic results are based on Equations 30 and 34; for simulations details, see Methods and Supplementary Section 6. The results correspond to setting $N = 2 \cdot 10^4$, $u = 10^{-8}$, per site per generation, $h^2 = 1/2$, and $C = 0.1$, and varying the target size L and threshold bias b_T as indicated. b) Prevalence vs the fraction of sites with large-effect sites in the model with 2 effect sizes. The model was solved as described in the text (see Supplementary Section 5 for details). The parameters are as in a, with $b_T = 1/2$ and $a_i/a_s = 100$. See Supplementary Fig. 3 for other choices of b_T and a_i/a_s . c) The impact of large-effect sites on the liability distribution. The 2 distributions shown correspond to the parameter values highlighted in panel b, with and without large-effect sites. The dotted outline shows a normal distribution with the same mean and variance as the Poisson convolution with large effects.

allele at a site is described by classic MSDB, with $E(x) \approx u/C$. Thus, in contrast to sites with large effect sizes at the lower end (and to sites with small effects), the selection acting on them and their genetic architecture are determined by the fitness cost of the disease; are insensitive to the relative position of the liability threshold, b_T ; and would be expected to quickly re-establish equilibrium after a change in population size.

Figure 4c shows the mapping between effect size and scaled selection coefficients for a large-effect site, in a model with 2 effect sizes. When the large effect size increases, the dependence of the strength of selection on model parameters varies gradually between the 2 behaviors that we described.

Disease prevalence

Lastly, we consider the disease prevalence at MSDB. The prevalence is equal to the area under the tail of the liability distribution that lies beyond the threshold, so calculating it requires knowledge of the shape of this distribution. The liability distribution is often assumed to be normal (see Discussion). This assumption seems sensible when the disease is highly polygenic and genetic contributions are small, such that an individual's liability arises from many small-effect genetic contributions and a normally distributed environmental contribution.

We begin by considering this case and assuming normality, which allows us to express the prevalence in terms of the density of the standardized liability distribution at the threshold. The standardized threshold density is

$$\varphi_T = f(T) \sqrt{V_P}, \quad (30)$$

where V_P is the (unstandardized) variance in liability, and the disease prevalence is

$$\bar{R} = 1 - \Phi(\varphi_+^{-1}(\varphi_T)) \quad (31)$$

where φ_+^{-1} is the positively defined inverse of the standard normal probability density function (PDF) and Φ is the standard normal

cumulative distribution function (CDF). As we assume that the disease is not exceedingly common or exceedingly rare (e.g. $10\% > \bar{R} > 0.1\%$), the dependence on the threshold density is approximately linear, with (Supplementary Fig. 2)

$$\bar{R} \approx 0.4 \cdot \varphi_T. \quad (32)$$

We can rely on our small-effect approximations to calculate the threshold density on the standard scale. Noting that in the small-effect approximation $\bar{\gamma} \approx 2NCf(T)\bar{a}$, noting the heritability of liability $h^2 = V_A/V_P$, and relying on our approximation for V_A (Equation 18), we find that

$$\varphi_T = f(T) \sqrt{V_A/h^2} \approx f(T) \sqrt{\frac{1}{h^2} \frac{2Lu \cdot \bar{a} b_T}{C \cdot f(T)}} = \left(\sqrt{Lu/(h^2 2N)/C} \right) \sqrt{\bar{\gamma} b_T}. \quad (33)$$

Under our assumption that effect sizes are small, the mean scaled selection coefficient $\bar{\gamma}$ and thus the term $\bar{\gamma} b_T$ are fully determined by the threshold bias, b_T , and distribution of effect sizes, g . In the single effect case, we can substitute the explicit expression for $\bar{\gamma}$ (Equation 26) to find that the standard threshold density

$$\varphi_T \approx \left(\sqrt{Lu/(h^2 2N)/C} \right) \sqrt{b_T \ln \left(\frac{1+b_T}{1-b_T} \right)}. \quad (34)$$

Figure 5a shows how the prevalence in the single effect size model increases with b_T for several possible values of the compound parameter $\sqrt{Lu/(h^2 2N)/C}$. As an illustration, we consider parameter values typical of humans, i.e. $N = 2 \cdot 10^4$ and $u = 10^{-8}$ per site per generation, a heritability $h^2 = 1/2$, and a cost $C = 0.1$. We vary the target size within the wide range estimated for complex, quantitative traits, e.g. $L = 1.5 \cdot 10^6$ to $1.5 \cdot 10^8$ (Simons et al. 2025). For these parameters, we find that a disease prevalence of 1% or greater at MSDB is attainable if the target size and/or the threshold bias is relatively large.

Next, we consider how sites with large effects impact prevalence. To this end, we study a model with 2 effect sizes: a fraction p_s of sites have a small effect size a_s and a fraction p_l of sites have a large effect size a_l , where $p_s + p_l = 1$. In this parametrization, the boundary case with $p_s = 1$ and $p_l = 0$ corresponds to the single effect model that we considered before, and changes in prevalence when we move away from this boundary by increasing p_l reflect the effect of increasing the proportion of large effects.

If we plausibly assume that individuals carry only a few large-effect risk-increasing alleles, then we no longer expect the liability distribution to be normal. Variation among individuals in the number of these alleles introduces a fat tail of individuals with higher liabilities and thus a skewed liability distribution. In the case with 2 effect sizes, the number of large-effect, risk-increasing alleles follows a Poisson distribution with mean $2Lp_l/s(a_l)$ (Felsenstein 1974), where we assume that the mean $2Lp_l/s(a_l) \leq 3$ so that an individual carries only at most a few of these alleles.

We therefore model the liability distribution at MSDB as arising from 3 components: (1) a large-effect genetic liability distribution that arises from the Poisson distributed number of large-effect risk-increasing alleles, (2) a normally distributed genetic liability arising from small-effect sites, and (3) a normally distributed environmental liability. The liability of an individual in the population is randomly sampled from the contributions to liability arising from each of these components. The liability distribution is therefore the convolution of these 3 distributions.

We solve this model numerically (for details, see Supplementary Section 5). We assume that all large-effect sites are fixed for the low-liability allele and derive the threshold density, $f(T)$, by requiring that the fixed bias from the small and large-effect sites combined matches the threshold position, b_T . Given the threshold density, we solve for the liability distribution arising from small-effect sites. We then solve for the selection coefficient of large-effect sites, $s(a_l)$, by requiring that the convolution of the normal and Poisson components of the liability distribution match the threshold density, $f(T)$.

Figure 5b shows the prevalence as a function of the proportion of large-effect sites. Here, we set $a_l/a_s = 100$ and $b_T = 1/2$. All other model parameters are set to the same values that we used for the case with a single effect size (as in Fig. 5a). We vary the proportion of large-effect sites between $p_l = 0$, corresponding to the case without large effects, and p_l such that $(2Lp_l/s(a_l)) = 3$. We validated our numerical solution against simulations (Supplementary Fig. 3).

Increasing the proportion of large-effect sites affects the liability distribution in 3 ways (Figs. 5c). First, it increases the variance in liability, because large-effect sites contribute much more variance than small-effect sites (Supplementary Section 2.3). Second, it introduces a right skew in liability due to the variation in the number of large-effect risk-increasing alleles. Third, it decreases the threshold density $f(T)$, because, with large-effect sites all fixed for the risk-decreasing alleles, the fixed bias and thus the selection at small-effect sites is weaker. We would expect the first 2 effects to increase the prevalence and the third one to decrease it. For the parameter values that we consider in Fig. 5b, the combination of these effects leads to a reduction in prevalence when the fraction of large-effect sites increases. In Supplementary Figs. 4–7, we consider how the prevalence changes with increasing fraction of large-effect sites for other choices of b_T and a_l/a_s . In general, we find that when b_T is closer to 0, increasing the fraction of large-effect sites tends to lead to a decrease in prevalence, whereas when it is closer to 1, it tends to lead to an increase.

Discussion

We introduced an evolutionary model of complex disease susceptibility, in which a variant's effect on disease risk follows from the liability threshold model and a variant's effect on fitness follows from its effect on disease risk. The model can be viewed as a generalization of the classic MSDB model of Mendelian diseases and is also closely related to models used to study genetic load (see Introduction). We solved the model for the genetic architecture and prevalence of the disease at MSDB.

MSDB in this model can be understood as a “mean field” equilibrium. Selection on sites with a given effect on liability is determined by the phenotypic distribution (the “field”), and the phenotypic distribution arises from the aggregate behavior at all sites (the “mean”). The density of the phenotypic distribution at the liability threshold is determined by matching the population's fixed state with the position of the threshold on the liability scale (given the distribution of liability effect sizes). The threshold density divides liability effect sizes into “small” and “large,” which differ in their mapping onto the effects on disease risk and fitness. For sites with small effects on the liability scale, the effects on disease risk and fitness are proportional to their liability effects and to the threshold density. For sites with large effects, the effects on disease risk and fitness depend on nonlinear cumulants of the liability distribution and the fitness cost of the disease. The selection acting on sites shapes their genetic architecture, which, together with environmental effects on liability, determines disease prevalence. We mapped out these relationships and their main implications for observable quantities and solved the model explicitly for simple distributions of effect sizes.

With these predictions in hand, we can ask whether our model fits what is known about the genetic architecture of complex disease susceptibility in humans. Figure 6 shows examples of the “smile” architecture of GWAS hits typical of complex diseases (Koch et al. 2024). These hits were ascertained in GWAS based on genotyping and imputation and therefore include common variants under weak and moderate selection but not variants under very strong selection. The “smile” architecture reflects selection in that variants with larger risk effects segregate at low minor allele frequencies (i.e. risk allele frequencies near 0 or 1). This is not the signature of selection expected under our model. Instead, we predict that if selection acted on variants due to their effects on disease risk, the architecture should be asymmetric, with a depletion of major alleles increasing risk, rather than (approximately) symmetric between minor and major risk-increasing alleles (compare Figs. 3b and 6).

Our results suggest that selection on variants due to their effects on disease risk should be stronger for diseases with greater fitness costs and higher prevalences (see, e.g. Equations 5 and 34). The departure from our predictions might seem all the more surprising then, given that complex diseases often entail substantial fitness costs and are quite common in contemporary human populations. Schizophrenia, for example, has a prevalence of 0.5% to 1% (Jablensky 2000; Saha et al. 2005; Chan et al. 2015; Simeone et al. 2015), affects several fitness components and has been estimated to reduce fertility by up to 75% (Haukka et al. 2003; Laursen and Munk-Olsen 2010; Power et al. 2013). Other diseases, such as type 2 diabetes, with a global prevalence of 6% (Ong et al. 2023), and multiple sclerosis, with a prevalence of ~0.3% in the United States (Nelson et al. 2019), are associated with a substantial increase in mortality rates and are plausibly associated with substantial reductions in fitness (Scalfari et al. 2013; Emerging Risk Factors Collaboration 2023; Graves et al. 2023).

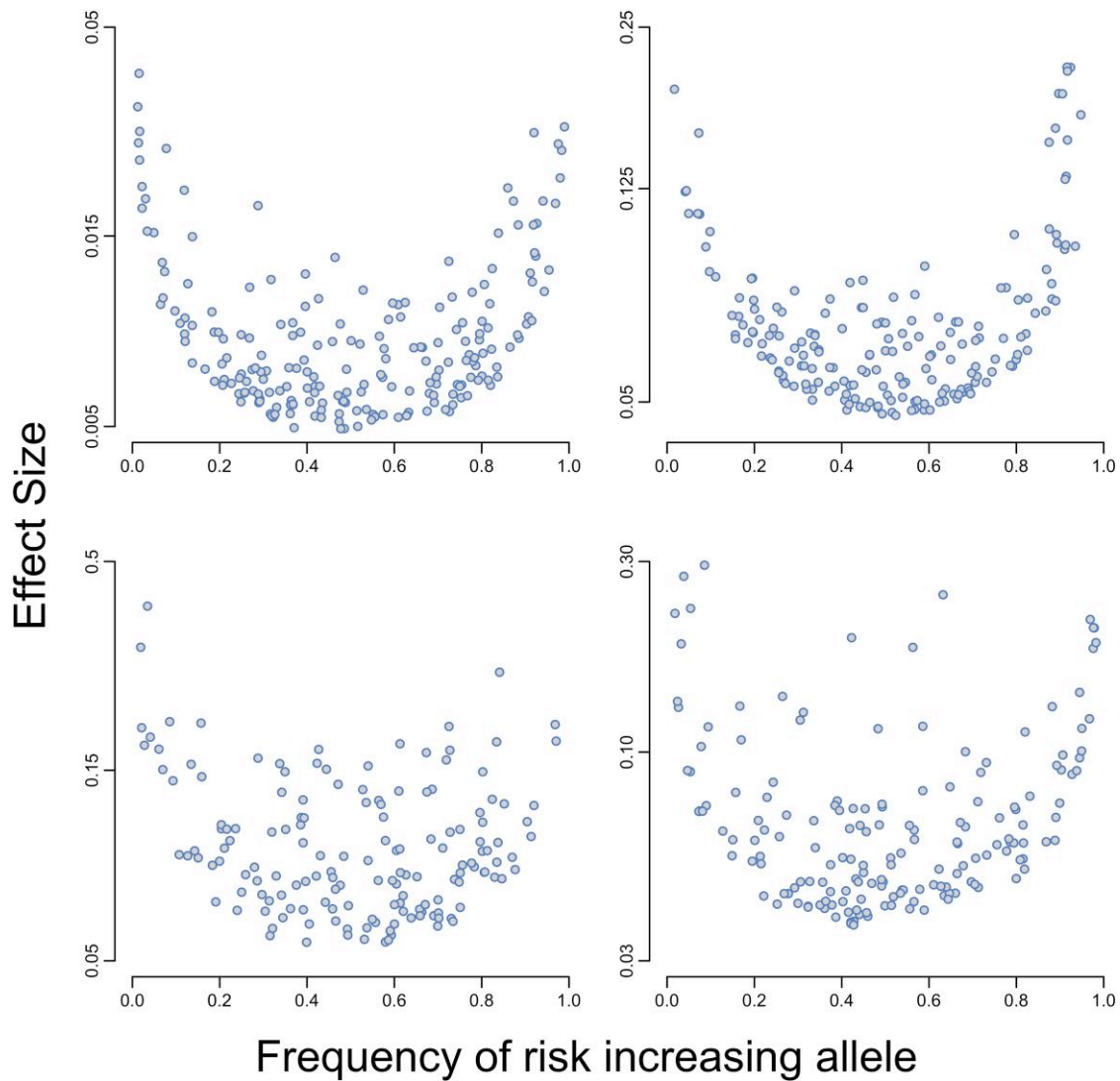


Fig. 6. The “smile” architecture of complex diseases in humans. Points correspond to approximately independent genome-wide significant hits. The data were taken from [Dönertaş et al. \(2021\)](#), [Trubetsky et al. \(2022\)](#), [Liu et al. \(2015\)](#), and [Spracklen et al. \(2020\)](#). See [Supplementary Section 7](#) for data processing details.

How then might we make sense of the fact that the architecture of common variation affecting complex disease susceptibility does not reflect selection against these diseases?

One possibility is that complex diseases that are common in contemporary human populations substantially increased in prevalence with very recent changes in environment. Examples plausibly include type 2 diabetes and other diseases associated with obesity ([Dai et al. 2020](#); [Teng et al. 2022](#); [Ong et al. 2023](#)), asthma ([Eder et al. 2006](#)), and autism ([Atladóttir et al. 2007](#); [Weintraub 2011](#); [Hansen et al. 2015](#)). More generally, we know little about the fitness cost and prevalence of human diseases more than a century back.

Persistent selection over molecular evolutionary timescales would be needed to attain the fixed bias that shapes the genetic architecture at MSDB in our model. These timescales vary with the scaled selection coefficients affecting sites. As an illustration, given the contemporary mutation rate in humans and assuming that scaled selection coefficients remain constant, sites with a scaled selection coefficient of $\gamma = 100$ would take on the order of a million generations to near MSDB (see, e.g. [Supplementary Section 2.2](#) in [Simons et al. 2014](#)); this roughly corresponds to 30

million years, extending back to the common ancestor of humans and Old World monkeys. With $\gamma = 1$, it would take on the order of 40 million generations. Disease biology and genetics has plausibly changed substantially over such timescales.

If a complex disease arose more recently, its genetic architecture would depend on the fixed bias that preceded its emergence. Even if this fixed state differed markedly from that at MSDB, we would expect genetic variation affecting disease to converge to a transient balance between the liability-increasing mutational pressure and liability-decreasing selection pressure within a population genetic timescale (on the order of N_e generations). This balance and the corresponding selection coefficients and genetic architecture would then be stable over population genetic timescales and should exhibit the kinds of insensitivity to changes in population size, environmental variance, and fitness cost of the disease that we described at MSDB. The genetic architecture, however, would crucially depend on the fixed state and thus on selection pressures that preceded the emergence of the disease, as well as on the subsequent pleiotropic selection pressures on genetic variation.

What could explain the “smile” architecture is if common variants affecting disease risk were selected because of their

pleiotropic effects on myriad quantitative traits that have been subject to stabilizing selection over long evolutionary timescales (see also Koch et al. 2024). The negative relationship between the effect sizes of the variants and minor allele frequencies would arise if the effects on diseases today are positively correlated with their effects on quantitative traits that were under stabilizing selection over the molecular evolutionary timescales that shape architecture at MSDB. The approximate symmetry between risk-increasing and risk-decreasing alleles would be expected if mutations with small and moderate fitness effects were (approximately) equally likely to increase or decrease disease risk (because new mutations are always selected against under long-term stabilizing selection). Together, these features would generate the “smile” architecture observed for many complex diseases.

This scenario seems plausible. The effects of most variants on complex traits are plausibly mediated by perturbations to the expression of genes in cellular, life history, and other contexts in which expression is held close to an optimal, nonzero value by stabilizing selection. Indeed, most heritable variance in complex traits arises from common regulatory variants (Yang et al. 2010; 2011; Finucane et al. 2015) in regions with open chromatin in the cellular contexts that affect these traits (Boyle et al. 2017; Sinnott-Armstrong et al. 2021; Spence et al. 2024). Moreover, there is an a priori expectation that genes that are expressed in a given context would have some nonzero, optimal expression level and evidence that selection generally acts against expression QTLs (eQTLs) (Mostafavi et al. 2023). In this genic perspective, larger perturbations to expression would have greater effects on traits and would be more strongly selected against (see, e.g. Conrad et al. 2006; Glassberg et al. 2019; Mostafavi et al. 2023; Zeng et al. 2024), leading to a negative relationship between effect sizes and minor allele frequencies. The approximate symmetry between trait-increasing and trait-decreasing alleles would arise if weakly and moderately selected perturbations to gene expression were approximately equally likely to increase or decrease gene expression.

The kind of pleiotropic stabilizing selection that would generate the “smile” architecture has been shown to explain key features of the genetic architecture observed in GWAS of highly polygenic quantitative traits. A simple model (with few parameters) of direct and pleiotropic stabilizing selection was shown to fit the joint distribution of frequencies and effect sizes of GWAS hits for highly polygenic quantitative traits in the UKB (Simons et al. 2025). A single parameter in the model describes the coupling between the effects of the variants on the trait and on fitness. The functional form of this relationship arises from assuming that genetic variation in the trait is highly pleiotropic (Simons et al. 2018). As in the genic case described above, this functional form associates larger effects on fitness with larger effects on a trait, and mutations affecting the trait are assumed to be equally likely to increase or decrease it, giving rise to a “smile” architecture. Additionally, the high polygenicity of complex diseases and quantitative traits that GWAS revealed has been partially attributed to “flattening”—whereby variants whose effects on a trait exceed a small threshold value have similar expected (asymptotic) contributions to variance in the trait (O’Connor et al. 2019). Such flattening arises under direct and pleiotropic stabilizing selection (Simons et al. 2018) but does not arise under the kind of directional selection modeled here (Supplementary Section 2.3).

We would expect there to be considerable overlap between common variation affecting complex quantitative traits and

complex diseases and consequently in the selection pressures that shape their genetic architecture. From a genic perspective, variation in gene expression in myriad contexts plausibly affects both. From the other end, quantitative traits like body mass index have been estimated to have mutational target sizes that exceed half of the fraction of the genome that has been estimated to be functional (Simons et al. 2025), and the high polygenicity of diseases like schizophrenia indicates similarly large target sizes (Loh et al. 2015). While the highly pleiotropic stabilizing selection model explains key observations about the genetic architecture of both complex quantitative traits and diseases, we cannot rule out there being alternative explanations for these observations. For example, a highly pleiotropic model of directional selection on traits in which the effects of variants on different traits are uncorrelated could potentially explain current observations (and could also be viewed as “apparent” stabilizing selection; Barton 1990; A. S. Kondrashov and Turelli 1992). Moreover, while the kind of directional selection we modeled falls short of explaining salient features of the architecture of common variation affecting complex disease susceptibility, we cannot rule out that it does have some effects on architecture.

Notably, the predictions of our model seem to be better aligned with the genetic architecture of rare variants with large effects on disease risk. These variants are too rare to be discovered in GWAS based on genotyping and imputation and were therefore discovered using other study designs, including association and burden tests based on whole-exome sequencing (Palmer et al. 2022; Singh et al. 2022) or whole-genome sequencing in quartet families (Satterstrom et al. 2020). In these studies, rare, large-effect alleles are generally found to increase disease risk as our model predicts. A caveat is that these studies have substantially greater power to identify rare risk-increasing alleles than rare risk-decreasing ones, so the observed asymmetry could also reflect an ascertainment bias. Nevertheless, the variants discovered in this way are often loss-of-function or copy number variants that appear to be under strong purifying selection, suggesting that the asymmetry is real.

This “large-effect” mode of architecture has been found for several common, complex diseases, notably autism spectrum disorder and schizophrenia (Satterstrom et al. 2020; Singh et al. 2022). Rare, strongly deleterious alleles are much younger than the common variation identified in GWAS and are therefore more likely to reflect selection on contemporary diseases. Purifying selection on these alleles could also reflect pleiotropic selection on other traits. However, alleles that are more specific in their effect on a given complex disease are expected to contribute more to heritable variance in that disease and are therefore also more likely to be identified in mapping studies (Spence et al. 2024).

The source of selection notwithstanding, complex disease architecture including both a strongly selected “large-effect” mode and a weakly selected “polygenic” mode should resemble the one that we modeled in having a fat-tailed liability distribution. Importantly, this architecture violates the normality often assumed in inference and theory (see e.g. Dempster and Lerner 1950; Falconer 1965). These departures from normality plausibly bias current estimates of the heritability of complex disease.

In particular, they would bias estimates of the proportional contributions of small and large-effect variants. Total liability-scale heritability is estimated from the correlations among relatives in disease state assuming: (i) that the liability distribution is Normal in order to derive the threshold density from

the prevalence (using our Equation 30), and (ii) that variant effect sizes are sufficiently small for liability effects to equal the ratio of risk effects and threshold density (using our Equation 8; Dempster and Lerner 1950; Falconer 1965). The contribution of small-effect variants to liability-scale heritability is estimated from GWAS based on the same assumptions (Hong Lee et al. 2011). When large-effect variants contribute substantially to heritability, the departure from these assumptions results in 2 main biases. First, the contribution of small-effect variants to the liability-scale heritability is underestimated (based on either GWAS or correlations among relatives) because assuming normality when the tail is fatter leads to an overestimate of the threshold density. Second, the relative contribution of large-effect variants to the total liability-scale heritability is overestimated because the linear transformation between risk and liability effects overestimates their contribution. Consequently, current estimates of the proportional contribution of large-effect variants are plausibly inflated. The magnitude of these biases depends on the departures from normality, which are unknown. These biases warrant further investigation.

In summary, there is compelling evidence that heritable variation in human complex traits, including in complex disease risk, evolves under MSDB (Sella and Barton 2019). Evidence from human GWAS and other study designs suggest that, at least for common genetic variation, the mode of selection in this balance is predominantly pleiotropic stabilizing selection (Simons et al. 2018; 2025; Koch et al. 2024; Spence et al. 2024). Rare genetic variation affecting complex disease susceptibility might also be shaped by directional selection of the kind we modeled here, and other modes of selection, such as balancing selection, doubtless contribute, but appear comparatively minor. More generally, as this work illustrates, by contrasting the predictions of evolutionary models with empirical findings, we can learn about the nature of selection affecting heritable variation in complex traits, and, more generally, about the evolutionary processes that shape inter-individual differences.

Data availability

Code for the simulations and numerical solutions of the model, as well as the scripts used to produce all figures can be downloaded at <https://github.com/jjberg2/msdbPaperCode>.

Supplemental material available at [GENETICS](https://www.genetics.org/online) online.

Acknowledgments

We thank Nick Barton, Magnus Nordborg, John Novembre, Molly Przeworski, and Himani Sachdeva for many helpful discussions and for comments on the manuscript, and we thank Joshua Schraiber and 2 anonymous reviewers for comments on the manuscript. We also thank members of the Sella, Przeworski and Andolfatto labs at Columbia University, and the Berg, Novembre and Steinrücken labs at the University of Chicago, for feedback on the work at various stages. This work was completed in part with resources provided by the University of Chicago's Research Computing Center.

Funding

This work was supported by National Institutes of Health F32 grant GM126787 and R35 grant GM151257 to J.J.B. and National Institutes of Health R01 grant GM115889 to G.S.

Conflicts of interest. We declare no conflicts of interest.

Literature cited

- Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 2023. 15 years of GWAS discovery: realizing the promise. *Am J Hum Genet.* 110: 179–194. <https://doi.org/10.1016/j.ajhg.2022.12.011>.
- Alon U. 2023. *Systems medicine*. 1st ed. Chapman and Hall/CRC.
- 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>.
- Atladóttir HÓ et al. 2007. Time trends in reported diagnoses of childhood neuropsychiatric disorders: a Danish cohort study. *Arch Pediatr Adolesc Med.* 161:193–198. <https://doi.org/10.1001/archpedi.161.2.193>.
- Barton NH. 1990. Pleiotropic models of quantitative variation. *Genetics.* 124:773–782. <https://doi.org/10.1093/genetics/124.3.773>.
- 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>.
- Chan KY et al. 2015. Prevalence of schizophrenia in China between 1990 and 2010. *J Glob Health.* 5:010410. <https://doi.org/10.7189/jogh.05.010410>.
- Charlesworth B. 1990. Mutation–selection balance and the evolutionary advantage of sex and recombination. *Genet Res (Camb).* 55:199–221. <https://doi.org/10.1017/S0016672300025532>.
- Charlesworth B. 2013a. Stabilizing selection, purifying selection, and mutational bias in finite populations. *Genetics.* 194:955–971. <https://doi.org/10.1534/genetics.113.151555>.
- Charlesworth B. 2013b. Why we are not dead one hundred times over. *Evolution.* 67:3354–3361. <https://doi.org/10.1111/evo.12195>.
- Clark AG. 1998. Mutation–selection balance with multiple alleles. In: Woodruff RC, Thompson JN, editors. *Mutation and evolution*. Springer Netherlands. p. 41–47. https://doi.org/10.1007/978-94-011-5210-5_4.
- Conrad DF, Andrews TD, Carter NP, Hurler ME, Pritchard JK. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet.* 38:75–81. <https://doi.org/10.1038/ng1697>.
- Crow JF. 1958. Some possibilities for measuring selection intensities in man. *Hum Biol.* 30:1–13. <https://www.jstor.org/stable/41449168>.
- Crow JF, Kimura M 1970. *An Introduction to population genetics theory*. The Blackburn Press.
- Dai H et al. 2020. The global burden of disease attributable to high body mass index in 195 countries and territories, 1990–2017: an analysis of the global burden of disease study. *PLoS Med.* 17: e1003198. <https://doi.org/10.1371/journal.pmed.1003198>.
- Danforth, Charles. 1923. The frequency of mutation and the incidence of hereditary traits in man. Paper presented at International Congress of Eugenics 1921, American Museum of Natural History. *Scientific Papers of the Second International Congress of Eugenics Held at American Museum of Natural History, New York, 1921 September 22–28*.
- Dempster ER, Lerner IM. 1950. Heritability of threshold characters. *Genetics.* 35:212–236. <https://doi.org/10.1093/genetics/35.2.212>.
- Dönertaş HM, Fabian DK, Fuentealba M, Partridge L, Thornton JM. 2021. Common genetic associations between age-related diseases. *Nat Aging.* 1:400–412. <https://doi.org/10.1038/s43587-021-00051-5>.
- Eder W, Ege MJ, von Mutius E. 2006. The asthma epidemic. *N Engl J Med.* 355:2226–2235. <https://doi.org/10.1056/NEJMra054308>.
- Emerging Risk Factors Collaboration. 2023. Life expectancy associated with different ages at diagnosis of type 2 diabetes in high-income countries: 23 million person-years of observation. *Lancet*

- Diabetes Endocrinol. 11:731–742. [https://doi.org/10.1016/S2213-8587\(23\)00223-1](https://doi.org/10.1016/S2213-8587(23)00223-1).
- Ewens W. 2004. *Mathematical population genetics 1: theoretical Introduction*. Springer.
- Falconer DS. 1965. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann Hum Genet.* 29: 51–76. <https://doi.org/10.1111/j.1469-1809.1965.tb00500.x>.
- Falconer DS, Mackay T. 1995. *Introduction to quantitative genetics*. Longman.
- Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics.* 78:737–756. <https://doi.org/10.1093/genetics/78.2.737>.
- Finucane HK et al. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet.* 47:1228–1235. <https://doi.org/10.1038/ng.3404>.
- Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. 2019. Measuring intolerance to mutation in human genetics. *Nat Genet.* 51:5. <https://doi.org/10.1038/s41588-019-0383-1>.
- Gillespie J. 2004. *Population genetics: a concise guide*. Johns Hopkins University Press.
- Glassberg EC, Gao Z, Harpak A, Lan X, Pritchard JK. 2019. Evidence for weak selective constraint on human gene expression. *Genetics.* 211:757–772. <https://doi.org/10.1534/genetics.118.301833>.
- Graves JS et al. 2023. Ageing and multiple sclerosis. *Lancet Neurol.* 22: 66–77. [https://doi.org/10.1016/S1474-4422\(22\)00184-3](https://doi.org/10.1016/S1474-4422(22)00184-3).
- Haldane JBS. 1927. A mathematical theory of natural and artificial selection, part V: selection and mutation. *Math Proceed Cambridge Phil 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. 2019. SLiM 3: forward genetic simulations beyond the wright–fisher model. *Mol Biol Evol.* 36:632–637. <https://doi.org/10.1093/molbev/msy228>.
- Hansen SN, Schendel DE, Parner ET. 2015. Explaining the increase in the prevalence of autism Spectrum disorders: the proportion attributable to changes in reporting practices. *JAMA Pediatr.* 169: 56–62. <https://doi.org/10.1001/jamapediatrics.2014.1893>.
- Harpak A, Bhaskar A, Pritchard JK. 2016. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. *PLoS Genet.* 12:e1006489. <https://doi.org/10.1371/journal.pgen.1006489>.
- Haukka J, Suvisaari J, Lönnqvist J. 2003. Fertility of patients with schizophrenia, their siblings, and the general population: a cohort study from 1950 to 1959 in Finland. *Am J Psychiatry.* 160: 460–463. <https://doi.org/10.1176/appi.ajp.160.3.460>.
- Hong Lee S, Wray NR, Goddard ME, Visscher PM. 2011. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet.* 88:294–305. <https://doi.org/10.1016/j.ajhg.2011.02.002>.
- Iwasa Y. 1988. Free fitness that always increases in evolution. *J Theor Biol.* 135:265–281. [https://doi.org/10.1016/S0022-5193\(88\)80243-1](https://doi.org/10.1016/S0022-5193(88)80243-1).
- Jablensky A. 2000. Epidemiology of schizophrenia: the global burden of disease and disability. *Eur Arch Psychiatry Clin Neurosci.* 250: 274–285. <https://doi.org/10.1007/s004060070002>.
- Jobling M, Hollox E, Hurles M, Kivisild T, Tyler-Smith C. 2013. *Human evolutionary genetics*. 2nd Edition Garland Science.
- Keightley PD, Hill WG. 1988. Quantitative genetic variability maintained by mutation-stabilizing selection balance in finite populations. *Genet Res (Camb).* 52:33–43. <https://doi.org/10.1017/S0016672300027282>.
- Kimura M, Crow JF. 1978. Effect of overall phenotypic selection on genetic change at individual loci. *Proceed Natl Acad Sci U S A.* 75:6168–6171. <https://doi.org/10.1073/pnas.75.12.6168>.
- Kimura M, Maruyama T. 1966. The mutational load with epistatic gene interactions in fitness. *Genetics.* 54:1337–1351. <https://doi.org/10.1093/genetics/54.6.1337>.
- Kimura M, Maruyama T, Crow JF. 1963. The mutation load in small populations. *Genetics.* 48:1303–1312. <https://doi.org/10.1093/genetics/48.10.1303>.
- King JL. 1966. The gene interaction component of the genetic load. *Genetics.* 53:403. <https://doi.org/10.1093/genetics/53.3.403>.
- Koch E, et al. 2024. Genetic association data are broadly consistent with stabilizing selection shaping human common diseases and traits. *bioRxiv.* <https://doi.org/10.1101/2024.06.19.599789>.
- Kondrashov AS. 1982. Selection against harmful mutations in large sexual and asexual populations. *Genet Res (Camb).* 40:325–332. <https://doi.org/10.1017/S0016672300019194>.
- Kondrashov AS. 1984. Deleterious mutations as an evolutionary factor: 1. The advantage of recombination. *Genet Res (Camb).* 44: 199–217. <https://doi.org/10.1017/S0016672300026392>.
- Kondrashov AS. 1988. Deleterious mutations and the evolution of sexual reproduction. *Nature.* 336:435–440. <https://doi.org/10.1038/336435a0>.
- Kondrashov AS. 1995. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *J Theor Biol.* 175:583–594. <https://doi.org/10.1006/jtbi.1995.0167>.
- Kondrashov AS. 2018. Through sex, nature is telling US something important. *Trends Genet.* 34:352–361. <https://doi.org/10.1016/j.tig.2018.01.003>.
- Kondrashov AS, Turelli M. 1992. Deleterious mutations, apparent stabilizing selection and the maintenance of quantitative variation." *investigations.* *Genetics.* 132:603–618. <https://doi.org/10.1093/genetics/132.2.603>.
- Lande R. 1975. The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genet Res (Camb).* 26: 221–235. <https://doi.org/10.1017/S0016672300016037>.
- Laursen TM, Munk-Olsen T. 2010. Reproductive patterns in psychotic patients. *Schizophr Res.* 121:234–240. <https://doi.org/10.1016/j.schres.2010.05.018>.
- Liu JZ et al. 2015. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet.* 47:979–986. <https://doi.org/10.1038/ng.3359>.
- Loh P-R et al. 2015. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet.* 47:1385–1392. <https://doi.org/10.1038/ng.3431>.
- Lush JL, Lamoreux WF, Hazel LN. 1948. The heritability of resistance to death in the fowl. *Poult Sci.* 27:375–388. <https://doi.org/10.3382/ps.0270375>.
- Milkman R. 1978. Selection differentials and selection coefficients. *Genetics.* 88:391–403. <https://doi.org/10.1093/genetics/88.2.391>.
- Mostafavi H, Spence JP, Naqvi S, Pritchard JK. 2023. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat Genet.* 55:1866–1875. <https://doi.org/10.1038/s41588-023-01529-1>.
- Muller HJ. 1950. Our load of mutations. *Am J Hum Genet.* 2:111–176. <https://doi.org/10.1093/ajhp/2.1.111>.
- Nelson LM et al. 2019. A new way to estimate neurologic disease prevalence in the United States. *Neurology.* 92:469–480. <https://doi.org/10.1212/WNL.0000000000007044>.
- O'Connor LJ et al. 2019. Extreme polygenicity of complex traits is explained by negative selection. *Am J Hum Genet.* 105:456–476. <https://doi.org/10.1016/j.ajhg.2019.07.003>.
- Ong KL et al. 2023. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a

- systematic analysis for the global burden of disease study 2021. *The Lancet*. 402:203–234. [https://doi.org/10.1016/S0140-6736\(23\)01301-6](https://doi.org/10.1016/S0140-6736(23)01301-6).
- Palmer DS et al. 2022. Exome sequencing in bipolar disorder identifies AKAP11 as a risk gene shared with schizophrenia. *Nat Genet*. 54: 541–547. <https://doi.org/10.1038/s41588-022-01034-x>.
- Power RA et al. 2013. Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia Nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry*. 70:22–30. <https://doi.org/10.1001/jamapsychiatry.2013.268>.
- Risch N. 1990. Linkage strategies for genetically Complex traits. I. Multilocus models. *Am J Hum Genet*. 46:222–228. PMC1684987.
- Robertson A. 1956. The effect of selection against extreme deviants based on deviation or on homozygosity. *J Genet*. 54:236. <https://doi.org/10.1007/BF02982779>.
- Saha S, Chant D, Welham J, McGrath J. 2005. A systematic review of the prevalence of schizophrenia. *PLoS Med*. 2:e141. <https://doi.org/10.1371/journal.pmed.0020141>.
- Satterstrom FK et al. 2020. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell*. 180:568–584.e23. <https://doi.org/10.1016/j.cell.2019.12.036>.
- Scalfari A et al. 2013. Mortality in patients with multiple sclerosis. *Neurology*. 81:184–192. <https://doi.org/10.1212/WNL.0b013e31829a3388>.
- Schoech AP et al. 2019. Quantification of frequency-dependent genetic architectures in 25 UK biobank traits reveals action of negative selection. *Nat Commun*. 10:790. <https://doi.org/10.1038/s41467-019-08424-6>.
- Schraiber JG, Spence JP, Edge MD. 2024. Estimation of demography and mutation rates from one million haploid genomes. *Am J Hum Genet*. 2152–2166. <https://doi.org/10.1101/2024.09.18.613708>.
- 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>.
- Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. *Proceed Natl Acad Sci*. 102:9541–9546. <https://doi.org/10.1073/pnas.0501865102>.
- Sham PC, Purcell SM. 2014. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet*. 15:335–346. <https://doi.org/10.1038/nrg3706>.
- Shi H, Kichaev G, Pasaniuc B. 2016. Contrasting the genetic architecture of 30 Complex traits from summary association data. *Am J Hum Genet*. 99:139–153. <https://doi.org/10.1016/j.ajhg.2016.05.013>.
- Simeone JC, Ward AJ, Rotella P, Collins J, Windisch R. 2015. An evaluation of variation in published estimates of schizophrenia prevalence from 1990–2013: a systematic literature review. *BMC Psychiatry*. 15:193. <https://doi.org/10.1186/s12888-015-0578-7>.
- 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>.
- Simons YB, Mostafavi H, Zhu Huisheng, Smith CJ, Pritchard JK, Sella G. 2025. Simple scaling laws control the genetic architectures of human complex traits. *PLoS Biol*. 23(10):e3003402. <https://doi.org/10.1371/journal.pbio.3003402>.
- Simons YB, Turchin MC, Pritchard JK, Sella G. 2014. The deleterious mutation load is insensitive to recent population history. *Nat Genet*. 46:220–224. <https://doi.org/10.1038/ng.2896>.
- Singh T et al. 2022. Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature*. 604:509–516. <https://doi.org/10.1038/s41586-022-04556-w>.
- Sinnott-Armstrong N, Naqvi S, Rivas M, Pritchard JK. 2021. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife*. 10:e58615. <https://doi.org/10.7554/eLife.58615>.
- Slatkin M. 2008. Exchangeable models of Complex inherited diseases. *Investigations. Genetics*. 179:2253–2261. <https://doi.org/10.1534/genetics.107.077719>.
- Spence JP et al. 2024 Dec 6. Specificity, length, and luck: how genes are prioritized by rare and common variant association studies. [preprint]. *bioRxiv*. <https://doi.org/10.1101/2024.12.12.628073>
- Spracklen CN et al. 2020. Identification of type 2 diabetes loci in 433,540 east Asian individuals. *Nature*. 582:240–245. <https://doi.org/10.1038/s41586-020-2263-3>.
- Teng ML et al. 2022. Global incidence and prevalence of nonalcoholic fatty liver disease. *Clin Mol Hepatol*. 29:S32. <https://doi.org/10.3350/cmh.2022.0365>.
- The Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 447:661–678. <https://doi.org/10.1038/nature05911>.
- Trubetskoy V et al. 2022. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*. 604:502–508. <https://doi.org/10.1038/s41586-022-04434-5>.
- Waxman D, Peck JR. 2003. The anomalous effects of biased mutation. *Genetics*. 164:1615–1626. <https://doi.org/10.1093/genetics/164.4.1615>.
- Weintraub K. 2011. The prevalence puzzle: autism counts. *Nature*. 479:22–24. <https://doi.org/10.1038/479022a>.
- Wray NR, Goddard ME. 2010. Multi-locus models of genetic risk of disease. *Genome Med*. 2:10. <https://doi.org/10.1186/gm131>.
- Wright S. 1926. A frequency curve adapted to variation in percentage occurrence. *J Am Stat Assoc*. 21:162. <https://doi.org/10.2307/2277143>.
- Wright S. 1934. An analysis of variability in number of digits in an inbred strain of Guinea pigs. *Genetics*. 19:506–536. <https://doi.org/10.1093/genetics/19.6.506>.
- Yang J et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 42:565–569. <https://doi.org/10.1038/ng.608>.
- Yang J et al. 2011. Genome partitioning of genetic variation for Complex traits using common SNPs. *Nat Genet*. 43:519–525. <https://doi.org/10.1038/ng.823>.
- Zeng J et al. 2021. Widespread signatures of natural selection across human Complex traits and functional genomic categories. *Nat Commun*. 12:1164. <https://doi.org/10.1038/s41467-021-21446-3>.
- Zeng JPS, Spence J, Mostafavi H, Pritchard JK. 2024. Bayesian estimation of gene constraint from an evolutionary model with gene features. *Nat Genet*. 56(8):1632–1643. <https://doi.org/10.1038/s41588-024-01820-9>.
- Zhang X-S, Hill WG. 2008. The anomalous effects of biased mutation revisited: mean–Optimum deviation and apparent directional selection under stabilizing selection. *Genetics*. 179:1135–1141. <https://doi.org/10.1534/genetics.107.083428>.