The Great Unanswered Question About Evolution

Dylan Maher

Drafted: 08Aug2023

> ...we are in the somewhat embarrassing position of observing some remarkably robust patterns...and yet seeing no compelling explanation for them. -Johnson and Barton (2005)

Author's Note: My treatment of the various models underlying the maintenance of quantitative variation closely follows the treatment given in the masterful Evolution and Selection of Quantitative Traits by Bruce Walsh and Michael Lynch (2018). All other sources are cited in-text.

1 Introduction

Without genetic variation, biological evolution would grind to a halt. A complete description of it therefore must be able to account for the various processes that maintain it—what gives rise to it, what depletes it, and how the flux between these plays out. While genetic variation can be influenced by migration and recombination, to keep things simple, we will be concerning ourselves here principally with selection, drift, and mutation.

1.1 Null Model

Anybody familiar with the basic framework of NHST probably knows our first step here: before we begin to model changes in variation, we model the lack of it. This might strike the uninitiated as counterintuitive at first, but the reasoning is fairly straightforward: if we want to be able to detect the presence of something, we had better know what the absence of it looks like. No point in trying to find a wet spot in the ocean. So what might be taking place to keep the the level of variation constant? Selection may or may not be taking place, so we set it aside (for now). Two things that will always be taking place however, are drift and mutation. The ultimate reasons for this are a bit abstract, but they boil down to the facts that populations are always of finite size and our evolutionary history has taken place in a changing environment. Thus we have our first class of models: mutation-drift models. What do these look like?

1.2 Lewontin's Paradox of Variation

Long before the days when "sequencing" (for the majority of geneticists) came to mean a procedure involving bubble wrap and a trip to FedEx, it was primarily a term used to refer to the analysis of proteins. The basic process was to take an organism, extract samples of its protein, and run it through a porous gel with electric current. The rate at which the protein moves through the gel provides an indication of its size. With some ampholyte and denaturing agents, you can even determine its charge. In the 1970s this technique became increasingly popular as a way for biologists to make inferences about protein variation (and by extension, genetic variation) within and among species. The method was straightforward and relatively inexpensive, leading to a cottage industry of research wherein geneticists (or probably more likely, field biology grad students) went out into nature, sampled a wide variety of organisms, ground them up to extract protein, and ran them out on gels (the so-called "find 'em and grind 'em" era of population genetics). One major conclusion from this research is what's known as Lewontin's Paradox of Variation.

1.3 Lewontin's Paradox of Variation

In a diploid (two alleles for each gene) population, where only drift and mutation are acting, we can quantify the amount of heterozygosity as a simple ratio of 2 times the mutation rate (μ) to the rate of drift (the reciprocal of 2 times the population size (N) yielding:

$$\frac{\text{rate of heterozygotes produced by mutation}}{\text{rate of heterozygotes lost by drift}} = \frac{2\mu}{\frac{1}{2N}} = 4N\mu$$

Making some simplifying assumptions, this is the expectation of nucleotide diversity (π), the average number of single nucleotide differences between haplotypes:

$$E(\pi) = \theta = 4N\mu$$

So assuming a constant mutation rate and population size, nucleotide diversity should scale in a more or less linear fashion. Luckily for us, we can do better than the protein analysis of yore and sequence genomes directly, giving us (with negligible error) an estimate of nucleotide diversity. As we can see, π shows only a

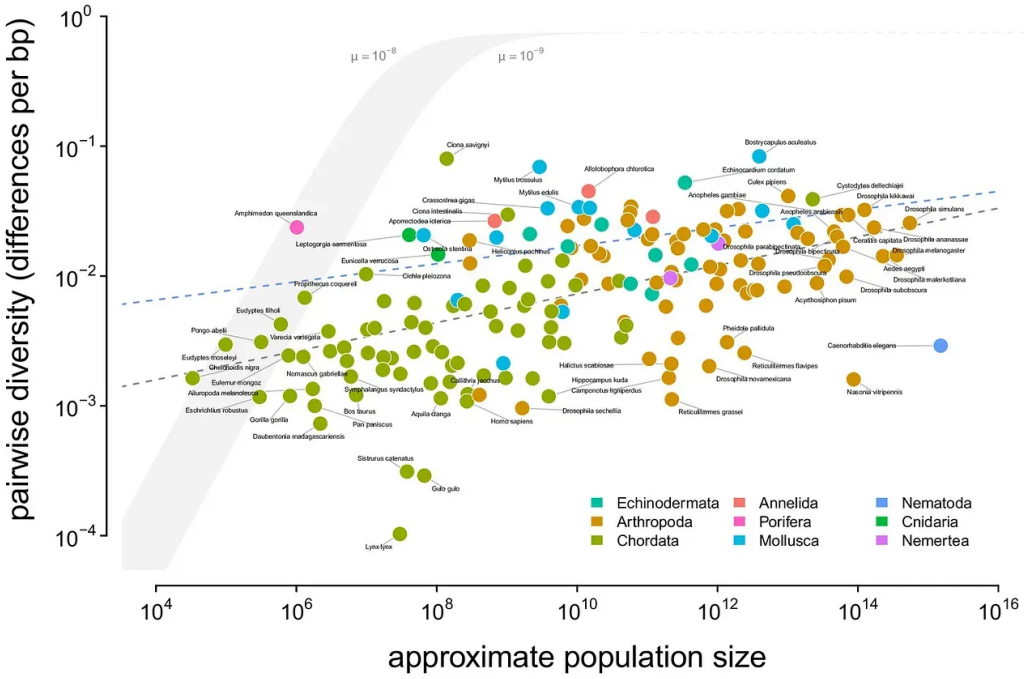


Figure 1: Pairwise diversity which varies over three orders of magnitude, shows a weak relationship with approximate population size, which varies over 12 orders of magnitude. Buffalo (2021)

weak relationship with population size. To be sure, there are possible explanations for this that are consistent with a mutation-drift model. For example, this pattern could be explained by mutation rates evolving to be correlated with population size. On the other hand, it is known that the levels of neutral (or nearly neutral) variation generally show a positive correlation with local recombination rate. This is an indirect result of the fact that alleles are not inherited singly, but rather by "chunks", formed through recombination. The lower the recombination rate, the larger the "chunk". If one of these chunks fixes, the larger it is, the less diversity there will be because all the polymorphisms (variation) in the chunk have "hitchhiked" together and are now fixed as well. Unlike population size however, we know this cannot be the result of a correlation with mutation rate as genomic regions with low levels of recombination do not show this same trend in between-species divergence. While debate continues regarding whether this is a product of recurrent sweeps or background selection, selection of some sort it must be. The takeaway here is that our mutation-drift model generates more variation than what we observe empirically, and we have good reason for thinking selection is pervasive. To have our model better accord with the data, we need to posit some action to decrease the level of variation, which brings us to selection.

2 Selection Models

We can incorporate the action of selection on a trait of interest broadly in one of two ways:

1. there is at least some direct selection acting on the trait
2. there is no direct selection acting on the trait, but the genetic variation giving rise to the trait has pleiotropic effects and some of those traits are under selection

2.1 Direct Selection Models

Direct selection models can be classified in four ways:

1. **Strict-Stabilizing Selection Models**
2. **Mutation-Stabilizing Selection Models**
3. **Balancing-Stabilizing Selection Models**
4. **Joint-Effects Models**

Under the strict-stabilizing selection model, we assume that stabilizing selection is acting on our trait of interest without the compensatory force of mutation. Under stabilizing selection, extreme phenotypes are selected against in favor of intermediate ones. This generates underdominance in fitness at the underlying loci and so iteration of this process in the absence of mutation necessarily leads an exhaustion of genetic variation—keep chopping the tails off a bell curve and eventually you will be left with a vertical line (variance = 0). You ("you" being here, a stand-in for life) might be able to survive a while in the absence of the degrading process of mutation but (heritable) adaptation to a changing environment is off the table.

So mutation and drift without selection leaves us with too much variation to be accounted for and and strict-stabilizing selection leaves us with none at all. Let's see how we do with mutation-stabilizing selection balance. As in the scenario above, extreme trait values are selected against, removing variation, but now new variation is introduced by mutation at the loci underlying the trait in question. Mutation-stabilizing selection models can be further specified by incorporating assumptions regarding the distribution of effect sizes of the mutations, leading to a class of models called continuum-of-allele (COA) models.

Continuum-of-allele models follow from relaxing the assumptions of Fisher's Infinitesimal Model. I don't think it's going too far to say Fisher's Infinitesimal Model essentially is the Modern Synthesis. When Darwin proposed his theory of natural selection, it caused quite a stir. Critics were legion and vocal. There were predictably, objections on religious grounds and likely some who just found the idea of sharing common ancestry with apes distatestful. But among the dissenters there was one who made a forceful (and accurate!) critique. Fleeming Jenkin, polyglot, polymath, and inventor of the cable car, pointed out there was a logical flaw in Darwin's theory. Rising to notoriety a mere six years before Mendel's Experiments on Plant Hybridization, Darwin was completely in the dark about the true mechanism of inheritance when he published The Origin of Species. Observing that offspring tended to have trait values roughly in between those of their parents, he suggested a mode of "blending inheritance," by which these offspring ended up as a sort of mix ("blend") of their parents. This issue with this is not dissimilar to the one we just finished discussing—if an offspring's traits are simply a sort of compromise between those of their parents and this process continues for long enough, any genetic variation giving rise to the trait will eventually be depleted. The resolution to this problem took over half a century and a genius on the order of Ronald Fisher. What Fisher proposed essentially was that what looked like blending inheritance was actually the infinitesimal effects of an (effectively) infinite number of segregating alleles. Random sampling of the parental alleles (through the processes of meiosis and recombination) then yields normally distributed phenotypes.

Obviously what is described by the infinitesimal model is an idealization. The extent of polygenicity as revealed by modern statistical genetics has proved quite a surprise for many, but as extreme as it may be, to consider it infinite can only ever be an approximation. Additionally, a complete understanding of evolutionary processes necessarily entails models that can account for phenotypes that span the entire range of monogenic to polygenic.

Selection can be modeled with respect to phenotype or genotype. On the phenotypic level, trait values of offspring can be modeled using the breeder's equation and Bulmer's equation (enabling predictions of the between-generation change in trait mean and variance, respectively). These predictions however, require the assumptions of a linear and homoscedastic parent-offspring regression. A wealth of empirical studies have shown that for a small number of generations, these approximations are quite reliable. On the genotypic level however, selection induces allele frequency changes, and because in reality, the number of alleles underlying a trait are not actually infinite, the approximations afforded by assuming that they are begin to break down.

As outlined by Turelli (2017), restrictions on the infinitesimal model can be considered as a nested hierarchy (Figure 2).

The broadest category here is the infinitesimal model itself: a "large" number of loci, each with vanishingly small effects. As allele frequency changes are a function of their effect size, vanishingly small effect produce vanishingly small frequency changes (small enough to be essentially ignored). Adding the restrictive assumption that within-family segregation variance depends only on the relatedness of the parents (and so is independent of parental phenotypes), yields the Gaussian descendants (or Fisher-Bulmer Infinitesimal) model and adding

Table 24.1 Classification of the different versions of the infinitesimal model, based on Turelli (2017). Note that these models are nested, such that a model makes all of the assumptions of any model that proceeded it in the table.

Infinitesimal genetics A large number of loci, each with vanishingly small effects.	Fisher (1918)
Gaussian descendants (Fisher-Bulmer Infinitesimal) Within-family segregation variance independent of parental phenotypes, depending only on the relatedness of parents.	Bossert (1963)
In the limit, results in a Gaussian distribution of breeding values in their unselected descendants. Parent-offspring regressions are linear and homoscedastic.	Fisher (1918), Bulmer (1971b), Barton et al. (2017)
Gaussian populations The distribution of breeding values in a population is Gaussian.	

Figure 2: Image from Walsh and Lynch (2018)

the further still more restrictive assumption that the distribution of breeding values in a population is Gaussian yields the Gaussian populations model. A complete treatment of COA models is outside the scope of this piece, but luckily for our purposes here, they happen to make qualitatively distinct predictions regarding the relative strength of mutation and selection.

Each of these models makes some assumption about the distribution of allelic effect size at a locus—specifically by making an assumption about a given allele's post-mutation effect size. In the Gaussian approximation, the post-mutation effect size is assumed to be the pre-mutation effect size plus some value drawn from a distribution with mean zero, while the House-of-Cards model assumes the post-mutation effect size and pre- mutation effect size are equivalent. Given the fact that the House-of-Cards model is indifferent to evolutionary history, simply drawing post-mutation effect sizes with an expected value of zero, whereas the Gaussian approximation is highly contingent on evolutionary history, assuming as it's "baseline" the existing allelic effect size, it follows that the cumulative effects of mutations will overpower selection in the former model, while being overpowered by selection in the latter. Each of these can be extended to incorporate the stochastic effects of drift (Stochastic House-of-Cards and Stochastic Gaussian, respectively), but their qualitative predictions remain the same.

The second way the removal of variation can be accounted for is by selectively favored pleiotropic fitness effects—that is, loci underlying a trait under stabilizing selection are also under balancing selection for another trait (the balancing-stabilizing selection model). The final class of models, joint-effects models, combines the previously separated aspects of mutation-stabilizing selection balance models and balancing- stabilizing selection models, so we have direct selection on the trait in question as well as 1) mutation on the underlying loci and 2) pleiotropic fitness effects.

If on the other hand, we would like to invoke the assumption that there is no direct selection on the trait in question, but instead strictly indirect selection from pleiotropic fitness effects, this leads us to pleiotropic models.

2.2 Pleiotropic Models

Note on terminology: Pleiotropic models assume the selection on the loci underlying the trait in question are due only to the action of selection on other traits, that are affected by the alleles. As a result, the trait in question is under no direct selection itself, so these models are sometimes referred to as neutral trait models. **Pleiotropic/Neutral Trait Models** can be classified in four ways:

- **Mutation-Drift Models**
- **Apparent Stabilizing Selection Models**
  - **Pleiotropic Overdominance Models**
  - **Pleiotropic Deleterious Mutation-Selection Balance Models**

As with direct selection models, our first modeling decision is whether or not to assume selection is taking place on the loci underlying the trait in question. If we assume it is not, this leads to mutation-drift models. (These

were the models that led to our discussion above about Lewontin's Paradox). If we assume it is, we are led to two classes of models that incorporate the scourge of empirical evolutionists everywhere: apparent (or spurious) stabilizing selection models: pleiotropic overdominant models and pleiotropic deleterious mutation-selection balance models.

The issue here is a subtle one and remains very much an open question. That is, how much of what appears to be stabilizing selection is *in fact* stabilizing selection. This is a natural extension of the complexities introduced by pleiotropy. It is entirely possible that when a trait appears to be under stabilizing selection, what is actually taking place is that stabilizing selection is acting on a trait whose loci just happen to have pleiotropic effects on both traits. The character of the obstacles to inference here naturally depend upon the traits in question: are their loci pleiotropic and if so, to what extent? As noted by Sella and Barton (2019), a closely related question is how polygenic the traits are. (emphasis added):

> The genetic basis of phenotypic variation spans a broad spectrum, from the Mendelian extreme, in which alleles at a single gene explain most heritable variance in a trait, to the highly polygenic extreme, in which thousands of alleles, distributed across the genome, contribute to heritable variance. **Strong evidence suggests that many traits fall in the latter extreme**... The high polygenicity and genome-wide distribution of variation of many traits are further supported by estimates of the heritable variation tagged by single-nucleotide polymorphisms (SNPs) in human GWASs (SNP heritability). For many traits, estimates of the heritability contributed by chromosomes are proportional to their length, suggesting that causal variants are numerous and are distributed fairly uniformly across the genome....**The high polygenicity of many traits almost inevitably implies extensive pleiotropy: If variation affecting a given trait spans a considerable portion of the functional genetic variation, then it is bound to overlap with variation affecting other traits (in addition to which causal variants in LD may also generate effective pleiotropy). Indeed, many of the variants identified in human GWASs are associated with more than one trait, and the extent of pleiotropy uncovered appears to be increasing rapidly with improvements in power and methodology.**

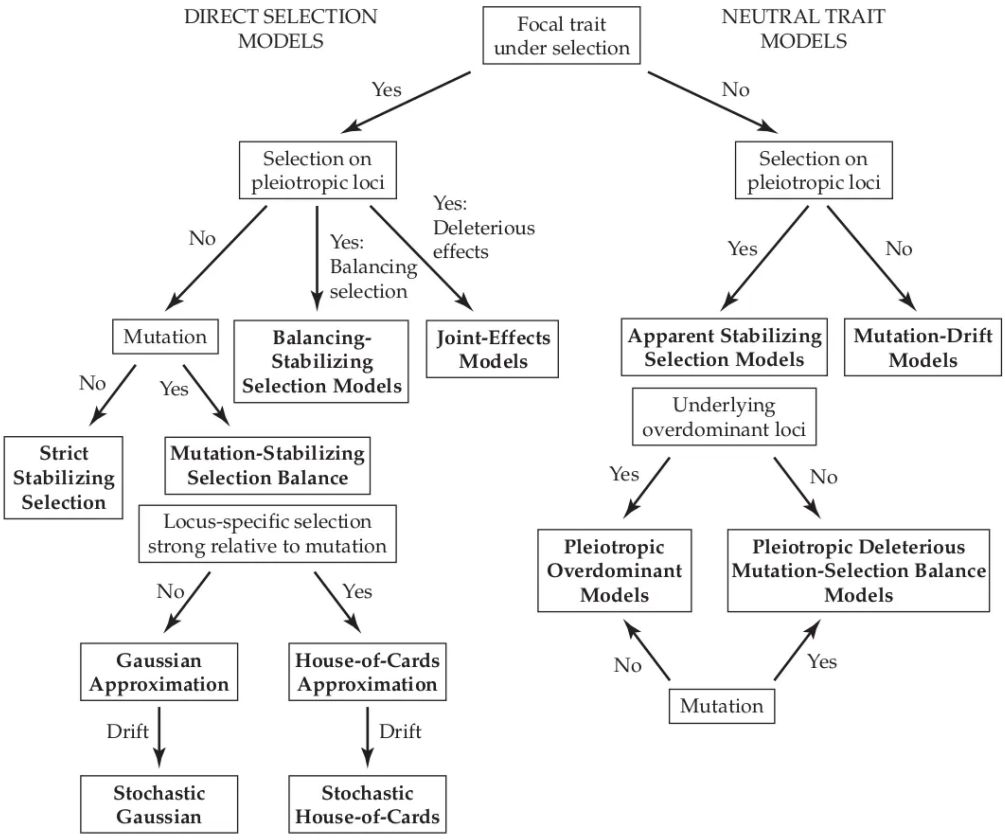


Figure 3: Image from Walsh and Lynch (2018)

3 Measurement Issues

So we have an array of models, all of which make some assumptions about the relative forces of mutation, drift, and selection—with this last having special emphasis placed on pleiotropy. As always, inference is contingent on measurement and in that respect, of all of these come with unique challenges. In the case of mutation, we have the classic statistical problem of inference about exceedingly rare events. This is complicated considerably by the facts that mutation rate varies not just between and within populations, but individuals (and even cells). It can also vary by type of mutation, biochemical environment, and chromosomal location. If I had to bet, I would say it likely can vary (systematically) in ways we are not even aware of. With the faster-than-Moore's Law decrease in price of sequencing, considerable progress has been made in this regard, but it should be borne in mind that these are still estimates and so come with estimation error. Selection may be the most problematic of

organism	mutations/ base pair/ replication	mutations/ base pair/ generation	mutations/ genome/ replication	BNID
multicellular				
human *H. sapiens*	10^{-10}	$1–4 \times 10^{-8}$ (mitochondria: 3×10^{-5})	0.2–1	105813, 100417, 105095, 108040, 109959, 105813, 110292, 111227, 111228
mouse *M. musculus*	2×10^{-10}	10^{-8}	0.5	100315, 106792, 100320
D. melanogaster	3×10^{-10}	10^{-8}	0.06	100365, 106793, 100370
C. elegans	$10^{-10}–10^{-9}$	10^{-8}	0.02–0.2	100290, 100287, 109959, 103520, 107886
unicellular				
bread mold *N. crassa*	10^{-10}		0.003	100355, 100359, 106747
budding yeast	$10^{-10}–10^{-9}$		0.003	100458, 100457, 109959, 110018
E. coli	$10^{-10}–10^{-9}$		0.0005–0.005	106748, 100269, 100263
DNA viruses				
bacteriophage T2 & T4	2×10^{-8}		0.004	103918, 103918
bacteriophage lambda	10^{-7}		0.004	100222, 105770, 100220
bacteriophage M13	10^{-6}		0.005	106788
RNA viruses				
bacteriophage Qβ	10^{-3}		7	106762
poliovirus	10^{-4}		1	106760
vesicular stomatitis virus	3×10^{-4}		4	106760
influenza A	10^{-5}		1	106760
RNA retroviruses				
spleen necrosis virus	2×10^{-5}		0.2	106762
moloney murine leukemia virus	4×10^{-6}		0.03	106760
rous sarcoma virus	5×10^{-5}		0.4	106762

Figure 4: Table from Milo and Phillips (2015)

the three. The major issue here is related to something we've already talked about. Selection does not respect the boundaries of neatly defined "traits" scientists are able (or inclined) to measure—it selects on phenotypes. These phenotypes can be decomposed (to an extent) into various traits but widespread correlations are bound to exist among them. What this means is that if we estimate selection on a given trait, what we are getting is both the direct effect of selection on that trait and the indirect effect of selection on any correlated traits. As mentioned, pleiotropy can induce correlations between traits but so too can linkage disequilibrium (LD). We haven't discussed LD much here but it's important to keep in mind that it's an important process behind the scenes. In addition to inducing trait correlation, it can be induced by selection, which in turn can depress effective population size, which in turn can increase the relative power of drift, which in turn can decrease the relative power of selection. You read that right.

While the extent of drift, being a (somewhat) straightforward function of population size, may seem the like the easiest of our forces to quantify, it too is not without its problems. I hinted at it above but when discussing Lewontin's Paradox, but I was not completely thorough about what is meant by "population size". In modeling evolution, what we are really concerned with is the effective population size. A detailed discussion of effective population size would take us pretty far afield but roughly, you can think of it as a measure of the degree to which drift can be expected to remove genetic variation from a population. The degree to which the effective population size differs from census population size (a "simple" count of all the organisms in a population) can be affected by a variety of things but some of the more salient are skewed reproduction rates between males

and females, changes in census population size, and variance in number of offspring between organisms. An important takeaway is that the effective population size is virtually always lower than the census population size and usually by quite a bit (think orders of magnitude).

Table 1 | **Effective population size (N_e) estimates from DNA sequence diversities**

Species	N_e	Genes used	Refs
Species with direct mutation rate estimates			
Humans	10,400	50 nuclear sequences	145
Drosophila melanogaster (African populations)	1,150,000	252 nuclear genes	108
Caenorhabditis elegans (self-fertilizing hermaphrodite)	80,000	6 nuclear genes	41
Escherichia coli	25,000,000	410 genes	146
Species with indirect mutation rate estimates			
Bonobo	12,300	50 nuclear sequences	145
Chimpanzee	21,300	50 nuclear sequences	145
Gorilla	25,200	50 nuclear sequences	145
Gray whale	34,410	9 nuclear gene introns	147
Caenorhabditis remanei (separate sexes)	1,600,000	6 nuclear genes	43
Plasmodium falciparum	210,000 –300,000	204 nuclear genes	148

For data from genes, synonymous site diversity for nuclear genes was used as the basis for the calculation, unless otherwise stated.

Figure 5: Table from Charlesworth (2009)

4 How Do the Models Hold Up?

Having outlined our various models and some of the relevant caveats, we can now ask how they hold up in light of empirical estimates. As a reminder, we have seven models altogether:

4.1 Direct Selection Models

Direct selection models can be classified in four ways:

1. **Strict-Stabilizing Selection Models**
2. **Mutation-Stabilizing Selection Models**
3. **Balancing-Stabilizing Selection Models**
4. **Joint-Effects Models**

- **Mutation-Drift Models**
- **Apparent Stabilizing Selection Models**
  - **Pleiotropic Overdominance Models**
  - **Pleiotropic Deleterious Mutation-Selection Balance Models**

4.2 Connecting Lewontin's Paradox to Phenotypic Variation

In our discussion of Lewontin's Paradox, we looked at the discrepancy between expected genetic variation as predicted by the neutral theory (mutation-drift models) and empirical estimates of genetic diversity. However at the time, we did not connect this in any way to phenotype, which is necessary since we will be relying on empirical estimates of parameters describing the extent of selection on phenotypes as well. To do this, we have to incorporate some function that links these predictions to phenotypic variation. A standard way of doing this is through the variance in phenotype arising in each generation as a result of mutation, called the mutational variance:

$$\sigma_m^2 = 2n\mu\sigma_\alpha^2$$

As before, $2n$ is the number of alleles, μ is the mutation rate, and σ_α^2 is the variance of mutational effects. The mutational variance is generally estimated by performing experiments with inbred lines—as these are genetically

identical, any difference that begins to appear can be inferred to be a result of mutation. This can then be scaled by the environmental variance to arrive at the mutational heritability (analogous to classic "heritability"—the variance in the population attributable to genetic differences between individuals):

$$h_m^2 = \frac{\sigma_m^2}{\sigma_E^2}$$

Under the assumption of additivity, the equilibirum level of mutational heritability is given by

$$\tilde{h}_m^2 = \frac{2N_e h_m^2}{1 + 2N_e h_m^2}$$

Table 12.1 Estimates of the mutational heritability for a variety of organisms and characters.

Species	Character	h_m^2	Reference
Drosophila melanogaster	Abdominal bristle number	0.0035	See text
	Sternopleural bristle number	0.0043	See text
	Enzyme activities	0.0022	Clark et al. 1995b Harada 1995
	Ethanol resistance	0.0009	Weber and Diggins 1990
	Body weight	0.0047	Clark et al. 1995b
	Wing dimensions	0.0020	Santiago et al. 1992
	Viability	0.0003	Mukai 1964 Mukai et al. 1972 Cardellino and Mukai 1975 Ohnishi 1977
Tribolium castaneum	Pupal weight	0.0091	Goodwill and Enfield 1971
Daphnia pulex	Life-history traits	0.0017	Lynch 1985
Mouse	Lengths of limb bones	0.0234	Bailey 1959
	Mandible measures	0.0231	Festing 1973
	Skull measures	0.0052	Carpenter et al. 1957 Deol et al. 1957 Hoi-Sen 1972
	6-week weight	0.0034	Caballero et al. 1995
Arabidopsis thaliana	Life-history traits	0.0039	Schultz et al. (in prep.)
Maize	Plant size	0.0112	Russell et al. 1963
	Reproductive traits	0.0073	Russell et al. 1963
Rice	Plant size	0.0030	Oka et al. 1958
	Reproductive traits	0.0028	Sakai and Suzuki 1964
Barley	Life-history traits	0.0002	Cox et al. 1987

Note: Where possible the results represent averages over multiple studies, ranging from analyses of drift among initially homozygous lines to artificial selection experiments. Detailed analyses of experiments performed prior to 1985 can be found in Lynch (1988b), and a more recent survey is contained in Houle et al. (1996).

Figure 6: Table from Lynch and Walsh (1998).

Let's try some back-of-the-envelope calculations with this. The estimates for *Drosophila melanogaster* (fruit flies) are fairly symmetric around the mean of 0.0026, so we'll start with that. The previous table (effective population sizes) gives an estimate for fruit flies of 1.15 million. This gives us a equilibrium level of mutational heritability of 0.9998328. Well. That seems... unlikely. Still, outliers exist. Let's see if we can get an estimate for something a little closer to home. You'll notice in the effective population size chart, humans are listed at 10,400. This might sound absurdly low but this is within the range used in that majority of the literature (typical estimates are between 10,000 and 20,000). The reason for this actually goes back to well before the Holocene—around the time the ancestors of all present-day non-Africans made their pilgrimage out of our home continent. As mentioned above, one of the most drastic ways to depress effective population size is to cull the

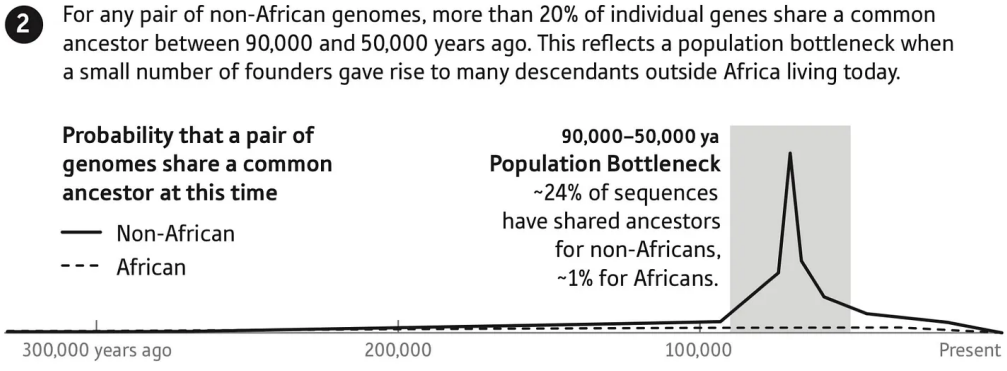
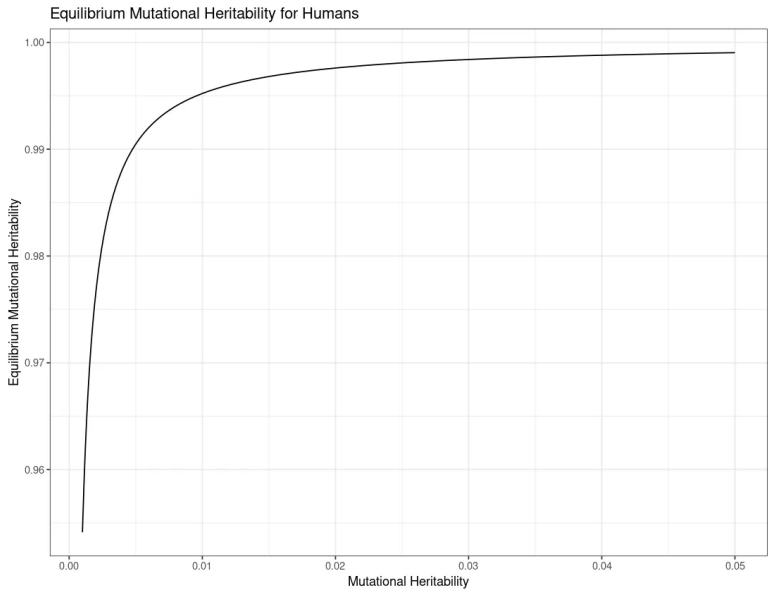


Figure 7: Image from Reich (2018).

census population down to a fraction of it's size. Current estimates have the Out-of-Africa expansion being on

the order of a 10-fold reduction, shrinking to an estimated 1,000-10,000 people, which likely remained at that size for quite some time.

IRBs are notoriously squeamish about performing inbred mutation accumulation experiments on humans, but since we only have one free parameter for our equilibrium value, we can see what it would be for a range of mutational heritabilities. Let's take 0.0001 as our lower bound (half that of barley, the lowest estimated value) and 0.05 as our upper bound (around twice that of morphological estimates in mice, the highest estimated values).



It appears our predicted equilibrium mutational heritability values are ≈ 0.95 at the lower end and asymptote at 1 rather quickly. This... also seems unlikely.

This is sort of a framing of the paradox of variation in phenotypic terms. On the genotypic level, under mutation-drift models, mutation introduces genetic variation faster than drift can remove it, so expected heterozygosity tends to its maximum. On the phenotypic level, mutation introduces genetic variation faster than than it introduces phenotypic variation (and faster than drift can remove it), so expected heritability reaches its maximum.

With this as a backdrop, we can now examine how our models look in light of estimated selection parameters.

We begin first with stabilizing selection-mutation models. We discussed above the effect of stabilizing selection on a single trait, so here we will broaden the scope to multiple traits. Our main parameters of interest will be the additive genetic variance σ_A^2, and the strength of selection, V_s. Fitness is commonly modeled as a Gaussian distribution with a variance of ω^2 and selection $s = \frac{1}{\omega^2}$. Tis means that higher values of V_s correspond to weaker selection and lower values of V_s correspond to stronger selection:

$$W(z) = e^{-(z-\theta)^2/\omega^2} \quad \text{with} \quad \omega^2 > 0$$

Walsh and Lynch cite Turelli's (1984) value of V_s as a typical estimate

$$V_s \approx 20\sigma_E^2 \implies \frac{V_s}{\sigma_E^2} \approx 20$$

As the narrow-sense heritability is given by

$$h^2 = \frac{\sigma_A^2}{\sigma_E^2}$$

we can take a reasonable estimate of heritability (0.5) and get

$$\frac{V_s}{\sigma_A^2} \approx 20$$

Now, assuming k independent traits, each under Gaussian selectin with common V_s, for populations at equilibrium, the reduction in mean population fitness for each trait is

$$\sqrt{\frac{V_s}{V_s + \sigma_A^2}}$$

Expansion by Taylor-series gives us an estimate of

$$\sqrt{\frac{V_s}{V_s + \sigma_A^2}} = \sqrt{\frac{1}{1 + \frac{\sigma_A^2}{V_s}}} \approx 1 - \frac{\sigma_A^2}{2V_s}$$

So for our k independent traits, this yields a cumulative fitness load of approximately

$$\exp\left(\frac{-k\sigma_A^2}{2V_s}\right)$$

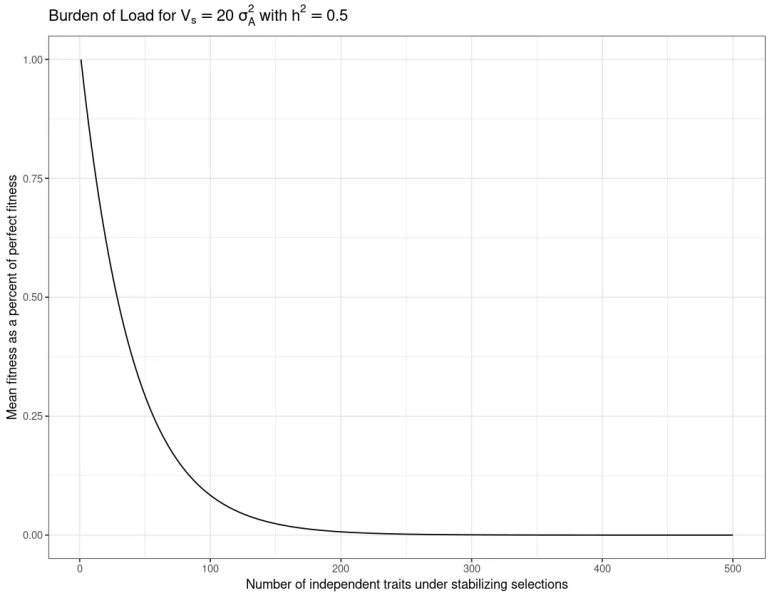
Assuming as above, a narrow-sense heritability of 0.5, we can then plug in the reciprocal of

$$\frac{V_s}{\sigma_A^2} \approx 20$$

to evaluate the fitness load on k traits as

$$\exp\left(\frac{-k}{40}\right)$$

As a function of k, this looks like



This is what is often referred to as the "cost of complexity." What it boils down to is that with high degrees of pleiotropy, the deleterious effects of mutations impose a fitness cost that scales as a function of the number of independent traits. Some intuition for this is provided by Fisher's Geometric Model.

5 Fisher's Geometric Model

Imagine two independent traits under stabilizing selection, governed by a joint Gaussian. A given phenotype (that is, a pair of trait values) is described by a point, z at distance d from the pair of trait values with the maximum possible fitness, θ. We let a random mutation, r be represented by a vector with a magnitude indicating its effect size. As mutations are random with respect to fitness, we suppose that it is just as likely to be extending out in any direction from its current value. If we were to hover over the this distribution, looking down on its peak, it would look something like Figure 8.

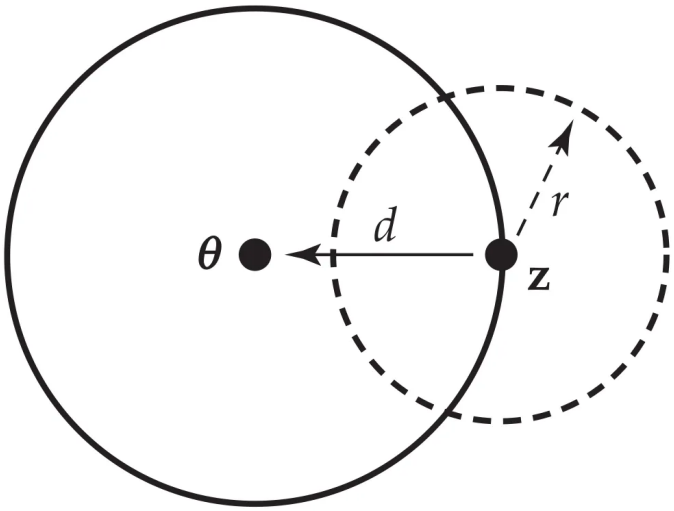


Figure 8: Image from Walsh and Lynch (2018).

Fisher's scaling parameter, x:

$$x = \frac{r\sqrt{n}}{2d}$$

is then sufficient to determine the probability of a random mutation being beneficial (i.e.moving our phenotype, z closer to θ):

$$P(\text{mutation is beneficial}) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(\frac{y}{2} dy\right) = 1 - \phi$$

where ϕ is the cdf of the standard normal distribution. Two key insights emerge from this. The first is that the probability of a mutation being beneficial is a function of two things: how far the current phenotype is from the optimum (d) and how large the effect size of the mutation is (r). This is easy to see by noticing the probability that a mutation is beneficial is equivalent to the proportion of the overlap between the circles governed by the radii of d and r (the area colored blue in Figure 9). The second insight is that as x is an increasing function

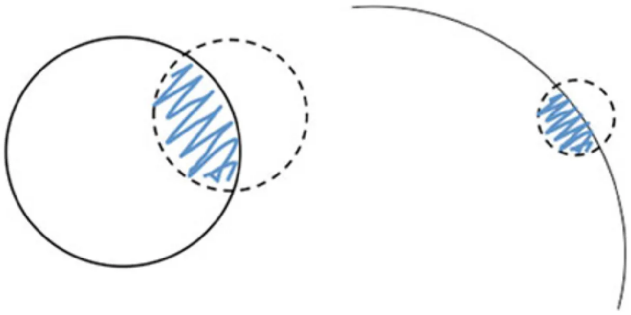


Figure 9: Image from Pavicel and Wagner (2012).

of the number of independent traits, the probability that a random mutation with a pleiotropic effect on these traits is beneficial decreases as the number of independent traits increases (Figure 10).

This is analogous to the situation with mean fitness, above. We have seen how stabilizing selection acting simultaneously on indepedent traits leads to a decrease in mean fitness as the number of independent traits increases and now we see this fitness load through the lens of the probability that mutations will be beneficial on independent traits.

So mutation-drift generates too much variation and strict-stabilizing selection removed it all. Incorporating mutation into stabilizing selection models seemed to be a step in the right direction but turns out to impose unsustainable fitness loads. Maybe these can be offset by a more forgiving fitness scheme?

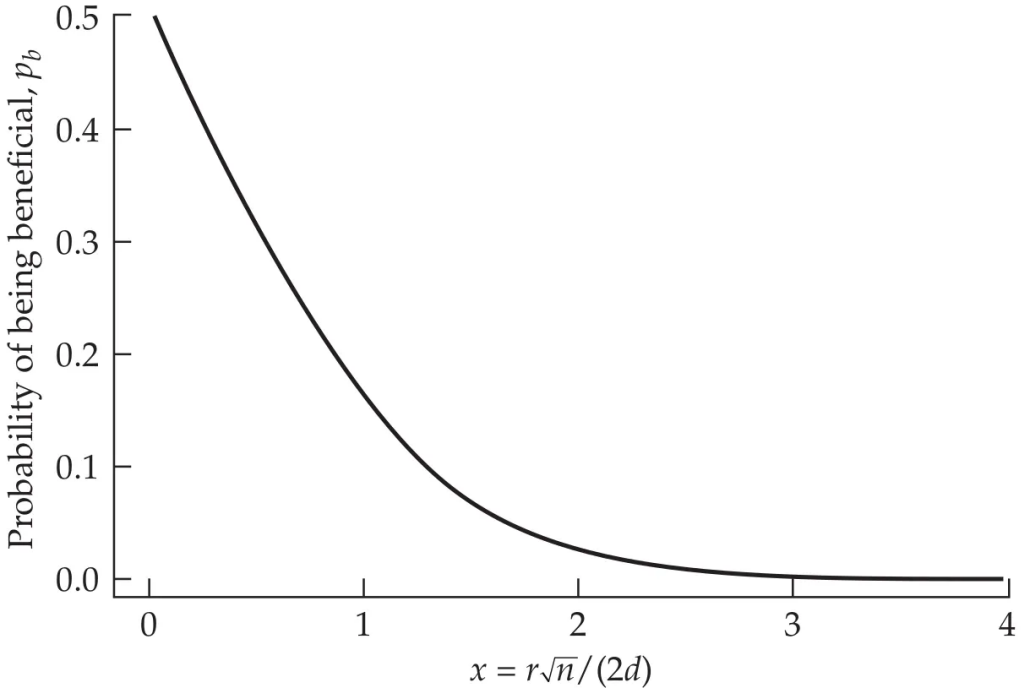


Figure 10: Image from Walsh and Lynch (2018).

5.1 Balancing-Stabilizing Selection Models

It's said that when trying to teach a subject, one should be very careful about incorporating exceptions. The reason for this is that exceptions tend to be surprising and counterintuitive, and so particularly memorable. Unfortunately, this is a lesson that genetics education has largely failed to learn. It's not surprising that the same schoolchildren who are taught about Mendel's Peas and Punnett squares grow up to have no trouble believing popular science articles about the discovery of the "gay gene" and other similarly biologically unlikely scenarios. The same I think, could be said for overdominance/heterozygote advantage. Read any standard genetics textbook and you will come across the "classic" example: HgbS, the allele responsible for sickle-cell anemia. Zero copies is the "wild-type" (most common in the population), two copies results in sickle-cell, but a goldilocks one copy confers resistance to malaria. In environments where malaria is endemic, this leads to the maintenance of the HgbS allele in the population, despite the fitness cost it imposes on homozygous mutants. What makes this example "classic" however is not that it is particularly informative or representative. In fact, in terms of understanding the broad strokes of evolutionary processes, there is good reason to think it is downright misleading. So what exactly is the evidence for heterozygote advantage? Well, it happens to be nicely summarized in Philip Hendrick's aptly named "What is the evidence for heterozygote advantage?" (2012):

> "Although there are some important examples of heterozygote advantage, overall the proportion of genes that appear to have polymorphism maintained by heterozygote advantage is small, based both on examination of particular candidate genes and on genomic surveys... Furthermore, except for the examples discussed here, heterozygote advantage (or even balancing selection) generally does not appear to be strong enough or act for long enough to maintain polymorphism over species."

As is always the case in biology, there are exceptions. Hendrick mentions a few: mutants in livestock and companion animals, warfarin resistance in rats, and the famous *ADH* (alcohol dehydrogenase) locus in *Drosophila*. There are also other types of balancing-selection. Negative frequency-dependent selection, for example. It seems likely this is what maintains variation in alleles at loci governing immune function—possibly owing to Red-Queen type arms races between genomes and various pathogens in the environment. Fluctuating selection and spacial variation are other potential candidates, but the conditions needed for these to maintain variation are often fairly restrictive. It should be said that like all types of selection, balancing-selection has its own unique difficulties when it comes to detection, which can introduce biases in the data we have available. Based on the current evidence however, it appears not to be a particularly compelling explanation for the maintenance of of genetic variation, generally. Unfortunately for us, this is also true for pleiotropic overdominance models, which additionally have problems of their own with fitness loads and maintaining sufficient variation. Fortunately, we have candidates models left in the wings. First up: pleiotropic deleterious mutation-selection balance.

5.2 Pleiotropic Deleterious Mutation-Selection Balance Models

First proposed by Hill and Keightley (1988), the HK model suggests that the additive variation can be maintained for a given (fitness neutral) trait as a side effect of largely deleterious pleiotropic loci that are in mutation-selection balance. The predictions their model makes can be described by four parameters: α, the effect of new mutations on focal traits, s, the effect of new mutations on fitness, $\tilde{p}$, the equilibrium frequency of allele under mutation-selection balance, μ, the mutation rate at allele's locus. Assuming negligible drift and an additive fitness model, the additive variance of this locus is given by the deceptively simple expression

$$2\alpha^2\tilde{p}(1-\tilde{p}) \approx 2\alpha^2\tilde{p} \approx \frac{2\alpha^2\mu}{s}$$

I say "deceptively simple" because in truth, both α and s are not constants, but random variables (i.e., described by probability distributions). Approximating the expectation of the above expression via Taylor expansion introduces (at minimum), three other terms: the kurtosis of mutational effects, the covariance between s and α^2 and the variance of s. This can be simplified a bit by assuming a large number of pleiotropic traits which each impact the locus with roughly similar selection coeffiecients. In this case, s can be approximated as a constant plus a bit of random error. Under these "constant-s" models, summing the above equation for n loci (assuming linkage equilibrium) gives the equilibrium value of the additive genetic variance as the ratio between the mutational variance and the selection coefficient:

$$\sigma_A^2 \approx \frac{\sigma_m^2}{s}$$

Using 10^{-3} as a reasonable estimate for s, we can rearrange this to get

$$\tilde{\sigma}_A^2 \approx \frac{\sigma_m^2}{s} \implies \sigma_m^2\tilde{\sigma}_A^2 s \implies \tilde{\sigma}_A^2(10^{-3})$$

Then, as in our *Drosophila* example from earlier, we can take 10^{-3} also as our mutational heritability to give us

$$h_m^2 = \frac{\sigma_m^2}{\sigma_E^2} \implies \sigma_m^2 = \sigma_E^2 h_m^2 \sigma_E^2(10^{-3})$$

implying, conveniently, an equilibrium heritability of 0.5:

$$\tilde{\sigma}_A^2 \approx_E^2 \implies \tilde{h}^2 \approx 0.5$$

So mutation-selection balance on mildly deleterious alleles can maintain observed levels of heritability (and so additive variance). This is good news! One of our recurrent problems so far has been an inability to simultaneously account for estimated selection pressure with observed levels of additive variance and here we have results that give some hope. We have to keep in mind the meaning of our terms, however. The constant-s model looks good for $s = 10^{-3}$, but in order to assess the plausibility of our assumed selection coefficient, we need to know what it implies at the phenotypic level—that is, for estimated strength of stabilizing selection. Skipping over quite a bit of math, what this model predicts about our strength of selection is

$$\hat{V}_s = \frac{\tilde{\sigma}_A^2}{s} = \frac{\sigma_E^2}{10^{-3}} \implies \hat{V}_s = 1000\sigma_E^2$$

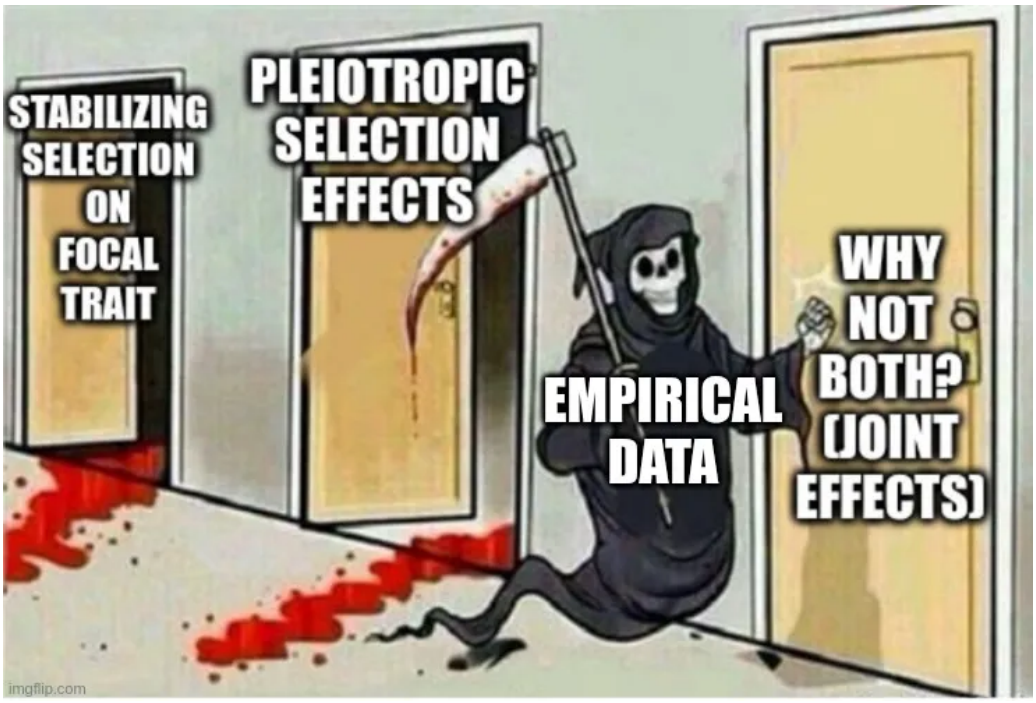
This is far weaker than observed levels of stabilizing selection, which as we have said, are generally closer to $20\sigma_E^2$. Using this observed term, our equilibrium heritability comes out to a whopping 0.02. So we see that although our constant-s model gives reasonable estimates at the genotypic level, we cannot square it with the degree of selection observed in nature. For those of you keeping score at home, this is where we stand:

1. ~~**Strict-Stabilizing Selection Models**~~
2. ~~**Mutation-Stabilizing Selection Models**~~
3. ~~**Balancing-Stabilizing Selection Models**~~
4. **Joint-Effects Models**

- ~~**Mutation-Drift Models**~~
- ~~**Apparent Stabilizing Selection Models**~~
  - ~~**Pleiotropic Overdominance Models**~~

– ~~Pleiotropic Deleterious Mutation-Selection Balance Models~~

Still the situation may not be so dire. We have ruled out strict direct selection models, which only makes sense given how pervasive pleiotropy is. We have also ruled out strict pleiotropic models, which also makes sense in its own way. As pervasive as pleiotropy is, it's likely that there is at least some direct selection taking place on any given trait. Acting on traits is selection's job, after all. Maybe there is hope yet?



6 Joint-Effects Models

So we have tried our hand at various degrees of parsimony and come up wanting. There is not much left to do except try the kitchen sink. This is the motivation behind joint- effects. The most extensive work done on these was a series of papers in the early to mid aughts by mathematical modeler Xu-Sheng Zhang and a renowned quantitative geneticist, the late William Hill. The models really do incorporate everything: direct stabilizing selection, pleiotropic stabilizing selection, mutation, even drift. In Appendix B of their 2002 paper they derive the equilibrium additive genetic variance predicted by their model as

$$\tilde{\sigma_A^2} \approx \sqrt{4n_s \times \sigma_m^2/\bar{s}_p}$$

where the new term, $\bar{s}_p$ is the mean pleiotropic selection coefficient. An interesting property of this genetic variance equilibrium value is that it happens to be the geometric mean between purely pleiotropic models and house-of-cards approximation models, which are

$$\tilde{\sigma}_A^2 = \frac{\sigma_E^2}{s}$$

and

$$\tilde{\sigma}_A^2 = 4V_s n\mu$$

respectively. So what does all this give us?

Taking h_m^2 as the typically assumed rate, 10^{-3} and assuming a 1:1 ratio for the variance of mutational effects to the mutational variance, the total mutation rate $2n\mu$ will be 0.02. Our strength of selection and selection coefficients will be the same as assumed earlier and we'll let the mean pleiotropic selection coefficient

be 0.005. So altogether we have

$$\begin{aligned} 2n\mu &= 0.02 \\ V_s &= 20\sigma_E^2 \\ \sigma_m^2 &= \frac{\sigma_E^2}{10^3} \\ \bar{s}_p &= 0.005 \end{aligned}$$

To find our predicted equilibrium genetic variance we take the geometric mean of the predicted values for the pure pleiotropy model and HCA:

$$\tilde{\sigma_A^2} = \frac{\sigma_m^2}{0.005} \quad \text{and} \quad \sigma_m^2 = \frac{\sigma_E^2}{10^3} \implies \tilde{\sigma_A^2} - 0.2\sigma_E^2$$

$$\tilde{\sigma}_A^2 = 4V_s n\mu = 4 * 20\sigma_E^2 * 0.01 = 0.8\sigma_E^2$$

$$\sqrt{0.2\sigma_E^2} \times 0.8\sigma_E^2 = 0.4\sigma_E^2$$

which implies a heritability of

$$h^2 = \frac{0.4}{1+0.4} \approx 0.29$$

Now we're cooking with gas! A heritability of 0.29 is on the lower side, but still within the range normally seen in natural populations. Before we get too excited, some caveats are in order. First and foremost, there were a lot of assumptions made along the way. One in particular is that an unfortunate property of the joint-effects model is if additive variance is high, the total strength of selection is approximately equal to the "true" strength of selection—meaning that spurious pleiotropy cannot cause much stronger selection than whatever the true selection is, which given high observed correlation between traits, seems fairly unlikely.

Further, our argument has assumed strictly additive effect for mutations. Unlike common alleles, where additivity is often a good approximation, mutations are expected to have serious dominance effects—lives are often defined by a single point mutation. Walsh and Lynch (2018) give their ruling on joint-effects models as follows:

> ...there appear to be regions of the parameter space under which the joint-effects model could account for both significant additive variation and sufficiently strong stabilizing selection. The unresolved issue is whether these regions are biologically realistic. There is also the secondary concern (from our previous load arguments) that strong direct stabilizing selection can only act on a limited number of traits, which suggest that weak true stabilizing selection is the norm, not the exception, significantly narrowing the size of these successful regions of the parameter space.

7 Dissenting Views and the Question of Evolvability

We have struck out. Zero out of seven of our proposed models seem to be able to account realistically and generally for the maintenance of genetic variation. In accordance with the measurement issues outlined above however, it is worth going into a bit more detail about some aspects of the data we have and what it implies. Additionally as I said up top, I have relied heavily on Walsh and Lynch's (2018) treatment of these issues and as such, I think it's only fair to give some alternate views a hearing. As an example of how parameters are interpreted, take the strength of selection we used above in our back-of-the-envelope joint-effects estimation (the assumption that the strength of selection is approximately twenty times the environmental variance). This estimate comes from a study by Michael Turelli in 1984. From Walsh and Lynch (2018):

> While Turelli's (1984) benchmark of $V_s \approx 20\sigma_E^2$ is typically assumed, the data today are both more extensive, and more problematic, than when he extracted this value from the literature... The relative constancy of many morphological phenotypes over evolutionary time is consistent with some form of stabilizing selection, as are the divergence data for gene-expression levels. However, the strength of such selection is far less clear. The meta-analysis by Kingsolver et al. (2001) on the quadratic term, γ, of a Lande-Arnold fitness gradient (Figure 30.5) shows that it is equally likely to be positive (disruptive selection) or negative (stabilizing selection). Conditioning on this value being negative, the mean strength is slightly stronger than Turelli's value ($\approx 10\sigma_E^2$). If correct, these higher estimates of V_s are more problematic for the previous models.

Walsh and Blows (2009) used these updated estimates from Kingsolver to make an argument similar the one we made in the discussion of Fisher's Geometric Theorem— about what pleiotropy and correlated selection imply for evolutionary potential (or more accurately, the lack thereof). Not everyone is convinced, however. Making

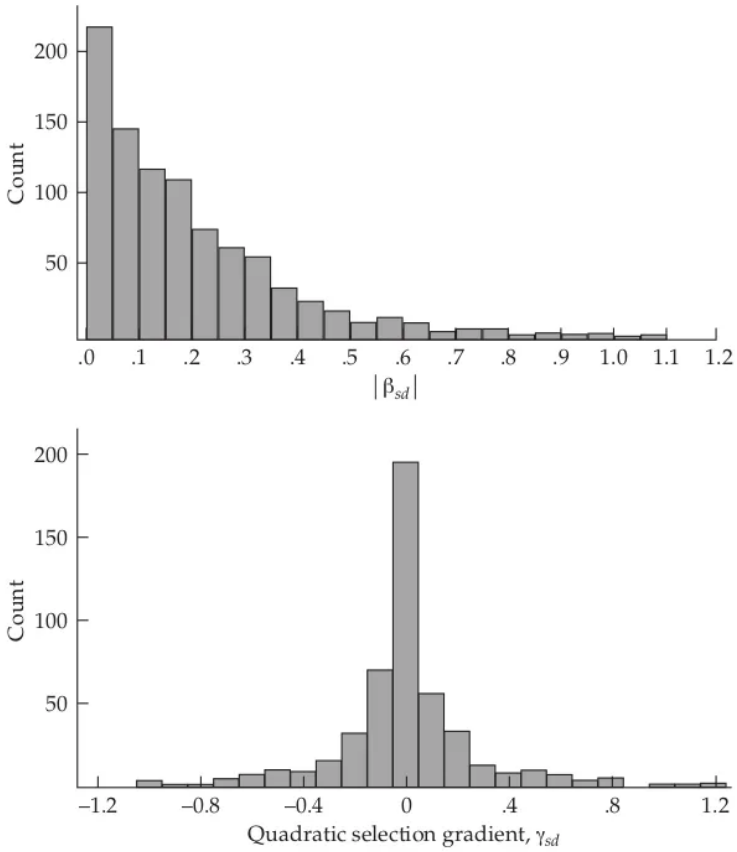


Figure 30.5 Summary of estimates of variance–standardized directional and quadratic gradients for natural populations. The data are from Kingsolver et al. (2001). **Top:** Plot of roughly 1000 estimated directional selection gradients in natural populations, with the median value of $|\beta_{sd}|$ being 0.16. The distribution of $|\beta_{sd}|$ was not significantly different from an exponential. **Bottom:** Plot of roughly 470 estimates of γ_{sd} from natural populations. The median value of $|\gamma_{sd}|$ was 0.10.

Figure 11: Figure 30.5, referenced above from Walsh and Lynch (2018). Originally from Kingsolver et al. (2001)

an argument which places emphasis on evolvability instead of heritability, Houle and Rossini (2022) argue that that instead of constraining evolutionary potential, "complexity represents an opportunity by increasing the evolutionary potential of a population." On the interpretation of Kingsolver's estimates they say

> Walsh and Blows (2009) offer an empirically based argument that absolute constraints become inevitable as the number of traits under selection increases. The linchpin of their argument is Johnson and Barton's (2005) conclusion that empirical estimates of stabilizing selection on typical traits are quite strong. . . .The key estimate of stabilizing selection is, however, based on a very curious interpretation of the complications of estimates of quadratic selection, γ, compiled by Kingsolver and colleagues (Kingsolver et al. 2012, 2001). The estimates of γ are widely distributed around a modal value of 0, but Johnson and Barton's estimate of the strength of stabilizing selection is the median of the negative estimates. . . .While the variance of the estimates is inflated by the substantial estimation error (Morrissey & Hadfield 2012), they are unlikely to be biased, suggesting that the strength of stabilizing is far weaker than Johnson and Barton's estimates, allowing far more traits to be subject to optimizing selection. As a result, the quantitative aspects of Walsh and Blows' (2009) argument are very much in doubt.

Hansen and Pélabon (2021), also advocating a more central role for evolvability in quantitative genetics suggest the role of epistasis has been in the main, incorrectly downplayed:

> It is remarkable that the rather obvious insight that biological epistasis influences the permanent response to selection has not made it into mainstream quantitative-genetics literature. Walsh and Lynch (2018), for example, devote an entire chapter to arguing that the effects of epistasis on the selection response are temporary and unimportant. This is mistaken. It has been shown analytically,

by simulations, and by experiment that epistasis can have permanent effects on the selection response (e.g. Carter et al. 2005, Hansen et al. 2006, Le Rouzic et al. 2008).

Putting forward their argument for evolvability, they say:

> The standard measure of evolutionary potential in EQG [evolutionary quantitative genetics] has been, and to some extent still is, the heritability. The justification for heritability as a measure of evolvability comes from the fact that heritability captures the fraction of phenotypic variance that is heritable and from its appearance in the breeder's equation. If evolvability is defined as the ability to respond to selection, and selection is measured as a selection differential, then evolvability is heritability.
>
> The problem with this logic has to do with correlations between the scale and the variables involved. If we assume the univariate Lande equation, and define evolvability, e, as response per strength of selection we get
>
> $$e \equiv \frac{\Delta \bar{z}}{\beta} = V_A e \equiv \frac{\Delta \bar{z}}{\beta} \equiv V_A$$
>
> If we now standardize all parts of the equation by the phenotypic standard deviation, we get
>
> $$e_\sigma \equiv \left(\frac{\Delta \bar{z}}{\sqrt{V_p}} \right) \left(\beta \sqrt{V_p} \right) = \frac{V_A}{V_P} \; h^2$$
>
> Where e_σ refers to variance standardization. Hence, the heritability, h^2 is a variance-standardized measure of evolvability. Alternatively, we can standardize by the trait mean to get
>
> $$e_\mu \equiv (\Delta \bar{z} / \bar{z}) = \frac{V_A}{\bar{z}^2} = I_A$$
>
> where I_A is the mean-standardized additive genetic variance. These choices of scale have dramatic effects. The h^2 and I_A are practiacally uncorrelated (Figure 1). They cannot both be good measures of the same thing (Hansen et al. 2011).

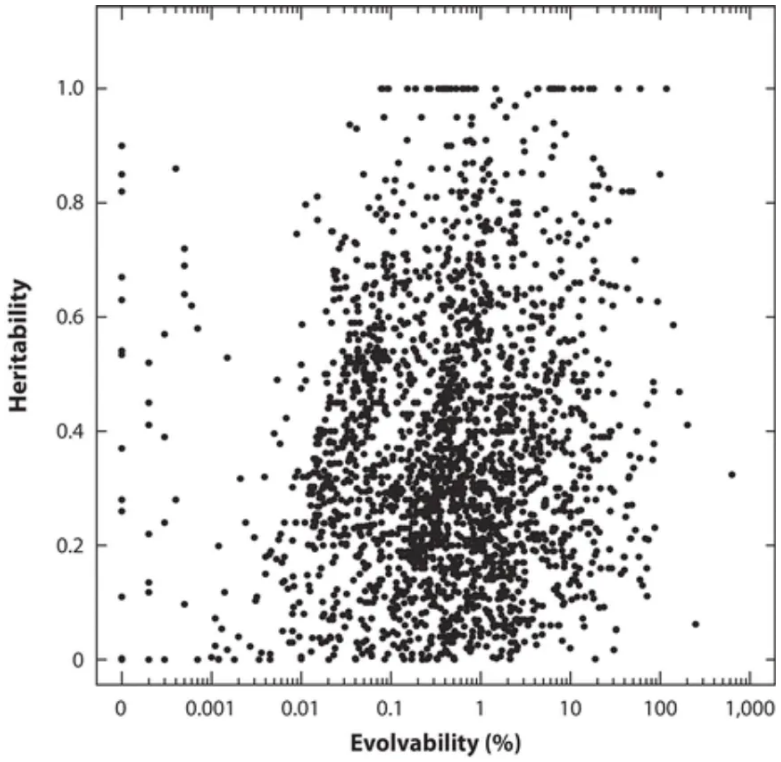


Figure 12: Figure 1, referenced in quote.

While the mathematical argument above is not in doubt, what the notion of evolvability does for us in terms of understanding the maintenance of genetic variation is less clear. Houle and Rossini contrast what they call the "complexity-as-cost" hypothesis in favor of a "complexity-as-opportunity" hypothesis. Their first move is

semantic: they characterize each of the hypotheses by recasting them terms of their definitions of "complexity" and "evolvability." In "complexity-as-cost," complexity is defined as the number of dimensions that can be simultaneously optimized by selection and evolvability is the probability of a new mutation being beneficial. (This latter is the function shown in the graph above in the FGM section). In "complexity-as-opportunity," complexity is defined as the size of the genome or number of traits subject to selection and evolvability is the set of variants subject to selection or drift. Their second move is to interrogate assumptions about adaptive landscapes, and what these imply.

8 Adaptive Landscapes

If you've thought at all about optimization problems, you probably have an intuitive idea about what adaptive landscapes are. Two classic models in evolutionary biology are Fisher's smooth adaptive landscape (consonant with his infinitesimal model and the properties conferred by the central limit theorem) and Wright's rugged adaptive landscape (in which he imagined his shifting-balance theory playing out). Pictured are examples

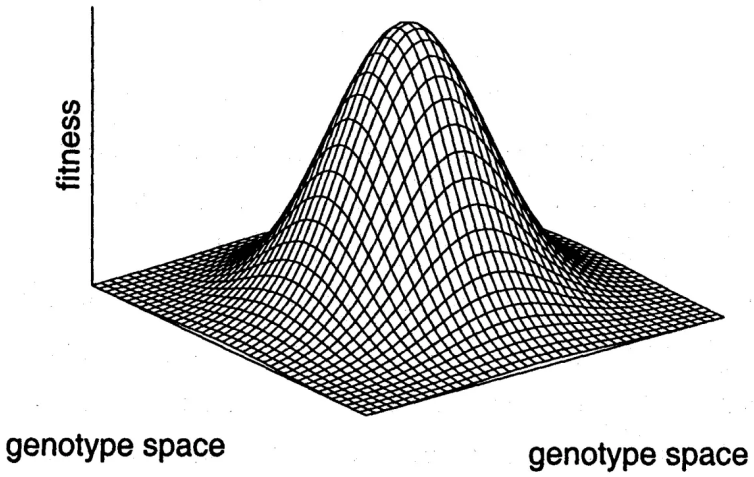


FIGURE 2.7. A single-peak fitness landscape.

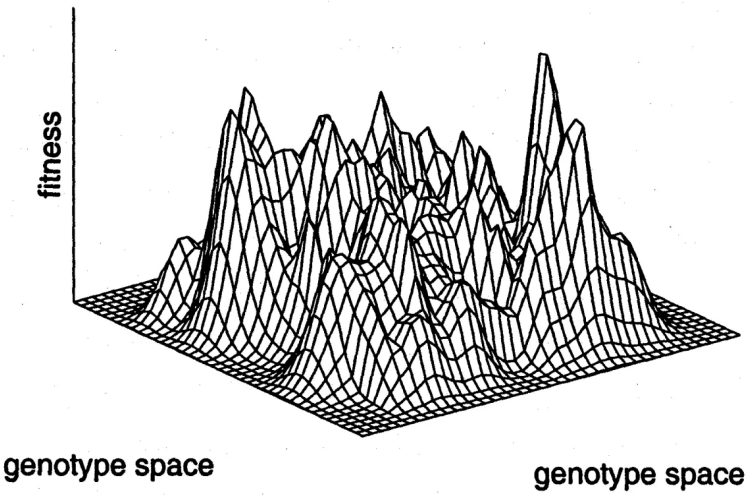


FIGURE 2.6. A rugged fitness landscape.

where fitness spans one dimension and genotype space, two—clearly extreme simplifications. (The dimensionality of a point is 0, the dimensionality of a line is 1, and the dimensionality of a plane is 2). The dimensionality of genotype space in the real world (?) is the number of "moves" available to a genotype. The human genome is ≈ 3.2 billion base pairs in length which makes ≈ 6.4 billion nucleotides. Four possible bases for each nucleotide means each starting nucleotide has three potential moves, which gives us a 19,200,000,000 dimensional space *for each starting genotype*. As the saying goes, if you find it difficult to picture a 19,200,000,000 dimensional

space, simply picture N-dimensional and then set $N := 19,200,000,000$.

Besides the dimensionality of genotype space, there is the topography to consider. With respect to allelic effects, this corresponds to the degree of epistasis—a smooth surface implies simple additivity while a rugged one implies interaction effects within or between loci. Houle and Rossini point out that while in a low-dimensional space, epistasis has a strong influence on the ruggedness of the landscape, in a high- dimensional landscape, this is no longer the case. They cite the work of Gavrilets (2004), who showed that because of this 1) in high-dimensional landscapes, populations were still able to "explore the breadth of G space, albeit on the slow timescale of drift" and 2) "the number of adjacent genotypes that are superior in fitness to the current genotype also increases with dimensionality."

Even with mind-bogglingly high-dimensionality, the models described above are still in truth, extreme simplifications. In reality, the adaptive landscape is not static, but fluid— fitness peaks are, especially in the long-term, moving targets. This means that rather than constantly hovering around "perfection," populations will frequently be in spaces where the constraints entailed by selection for various traits simultaneously are dramatically reduced.

A thread running throughout their argument is the idea of modularity. According to the complexity-as-cost view, modularity is what you would naturally expect to see as a result of the cipher of selection on traits that can vary in their degree of correlation and form clusters. Houle and Rossini argue "The assumption that integration and modularity align with the pattern of natural selection, and therefore help to explain the pattern of evolution, is largely untested. We have essentially no information on the patterns of natural selection to compare with data on variation within or among populations." Granting their point, it is not difficult to imagine why this is the case. Sussing out the patterns of correlated selection is difficult enough, to say nothing about its mechanisms. Despite the opportunities they posit as potentially afforded by complexity they close on a fairly agnostic note:

> The conviction that complexity affects evolvability positively or negatively is not supported by solid evidence. The answer is likely to be highly contingent on exactly which traits are considered, and most importantly, on the nature of natural selection. We lack good information about the nature of selection and whether organisms are generally at fitness peaks or wandering in a sea of shifting fitnesses. The harder we look at these issues, the less certain we are of how to think about evolvability in relation to complexity.

9 Conclusion

...