2.7 Natural Selection III. Genome-wide extent of selection

We have now touched on the main types of natural selection, and I have already hinted at a key question: how important are each of these in practice? Here we tackle this key question.

First though, it's helpful to give some historical context [310].

By the 1960s, much of the basic theory of population genetics had already been developed, but molecular techniques for measuring genetic variation were extremely limited. Consequently, population genetics was largely a theoretical field [311]. Little was known about the relative importance of the fundamental processes: mutation, recombination, migration, and drift; negative selection, positive selection, and balancing selection.

This started to change with the invention of **gel electrophoresis**, a technique that made it possible to measure protein variation on gels [312]. The first examples came from humans and flies in 1966 [313]. These first studies were followed by a flurry of electrophoresis studies in a wide range of organisms – so many that this was cheekily referred to as the "find 'em and grind 'em" approach [314].

Before the electrophoresis era it was anticipated that most protein variants would be subject to strong selection. Thus the default state would be a wildtype allele and perhaps additional rare deleterious variants; meanwhile there would be occasional rapid sweeps, and perhaps balancing selection in some genes [315].

Given these expectations, it was a surprise to find that protein variation is widespread in most species. For example, in 1966 Lewontin and Hubby estimated that around $^1/_4$ to $^1/_3$ of genes were polymorphic within populations of the fly Drosophila pseudoobscura [316]. One possibility was that this might indicate huge amounts of balancing selection, but this conclusion was controversial.

A complementary insight came from emerging data on protein differences between species. By the mid-1960s it was becoming apparent that proteins tend to accumulate amino acid substitutions steadily over evolutionary time. This was referred to by Zuckerkandl and Pauling in 1965 as the **molecular clock** [317]. One vivid illustration of the molecular clock was published by Richard Dickerson, below, in 1971 [318]:

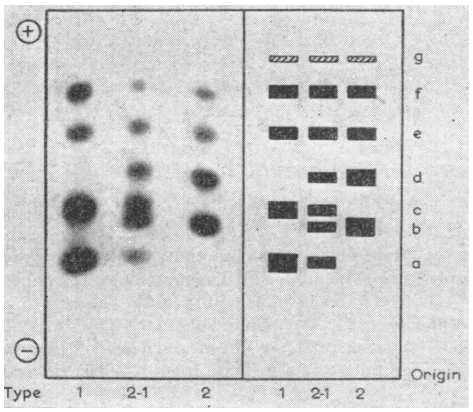

Figure 2.114: **Enzyme Polymorphisms in Man (1966).** *Gel electrophoresis, as shown here, made it possible to survey genetic variation for the first time.* *The vertical lanes show banding patterns for two alleles at the phosphoglucomutase enzyme: homozygotes in lanes 1, and 2; heterozygotes are a mixture of both patterns: 2-1. Experimental data at left, and a schematic of the banding patterns at right. The alleles were reported at frequencies 0.75 and 0.25 respectively, in human populations. Credit: Fig. 68 from Harry Harris (1966). [Link] Used with permission.*

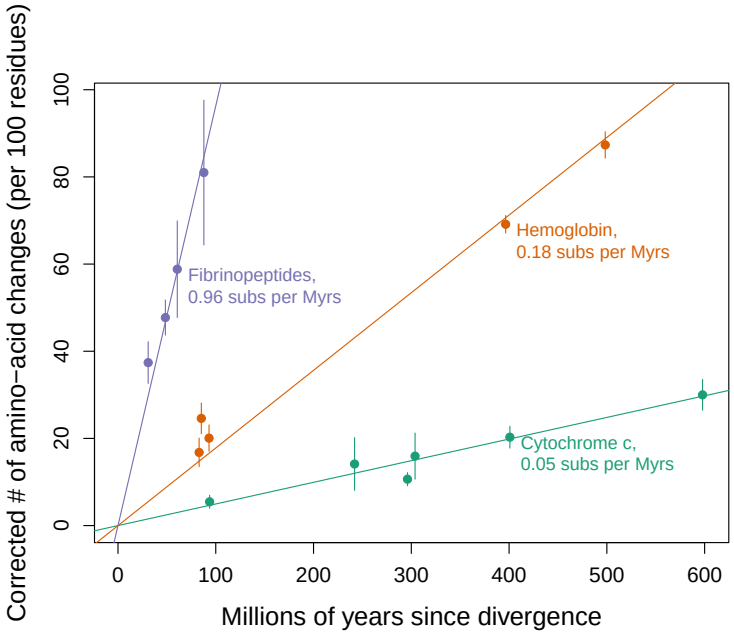

Figure 2.115: **One of the first demonstrations of the molecular clock, from 1971.** *The x-axis shows divergence times of pairs of species, as estimated from the fossil record; the y-axis shows fractions of amino acid differences in three proteins. The analysis was important for showing that protein differences accumulate roughly linearly over evolutionary time, but at different rates for different proteins.* *Credit: Figure 5.3 from Graham Coop in Population and Quantitative Genetics [Link], CC BY 3.0; based on Dickerson (1971).*

Of course it was possible that these protein changes were all adaptive, but even in the 1960s there were reasons to doubt this. The gene Cytochrome C, shown above, is found in a wide range of eukaryotes and fulfills a conserved role in the electron transport system in the mitochondria. King and Jukes (1968) noted that experiments comparing Cytochrome C proteins from different species could detect no functional differences. They hypothesized that the observed substitutions are mainly at positions that *do not* have a functional impact, and have fixed by neutral drift [319]. Their proposal contrasts sharply with an adaptive model of protein evolution, where one might expect most substitutions to be functional.

The Neutral Theory of Molecular Evolution. Together, these observations stimulated a paradigm shift in the late 1970s in how people thought about the main forces acting on genetic variation – and especially the role of genetic drift. These new ideas were articulated in particular by the Japanese scientist Motoo Kimura, who dubbed this the Neutral Theory of Molecular Evolution [320]. **In short, he proposed that most new mutations are either approximately neutral, or deleterious; advantageous mutations are very rare and contribute only a tiny fraction of polymorphism and differences between species.**

As stated by Kimura (1983) [321]: *"The neutral theory asserts that a great majority of evolutionary changes at the molecular level...are caused not by Darwinian selection but by random drift of selective neutral or nearly neutral mutants.... (P)olymorphisms are mainly due to mutations that are nearly enough neutral... that their behavior and fate are mainly determined by mutation and random drift...*

On the topic of selection he clarified that: *"The theory does not ... assume that selection plays no role; however, it does deny that any appreciable fraction of molecular change is due to positive selection or that molecular polymorphisms are determined by balanced selective forces... selective constraints imposed by negative selection are a very important part of the neutralist explanation..."*

It's hard to overstate the impact this model has had on how we think about genetic variation. The Neutral Theory provides an intellectual framework for thinking about modeling, and a null hypothesis for data analysis. It is no longer controversial that most new mutations are neutral and that, of those that are not neutral, most are selected against. As we'll discuss in Part 3 of the book, these properties allow us to use genetic variation as a tool for studying population structure and history while largely ignoring the role of selection.

That said, it's worth noting that early conceptions of the Neutral Theory under-appreciated the importance of some processes in shaping patterns of variation [322]. One important early addition came from Tomoko Ohta's work emphasizing the importance of **nearly neutral mutations**. Starting in 1973, Ohta argued that many mutations may have selection coefficients that are close to, but not precisely, 0. These can have important consequences: for example, recalling that selection is ineffective when $|4Ns|$ is less than about 1, we see that weakly deleterious variants fix at a higher rate in species with small N than in species with large N, which can affect substitution rates in different lineages (Chapter 2.5) [323].

Another under-appreciated area was the role of **linked selection** (which we'll cover in this chapter), and a third is the role of **polygenic stabilizing selection and adaptation** (Chapter 2.7).

The original theory also predated modern understandings of genome architecture, as well as the central importance of **gene regulation** in phenotypic variation and evolution.

And despite the Neutral's Theory's importance as a null hypothesis, significant effort in the last 50 years has been devoted to understanding its limitations. There has been a great deal of work aimed both at measuring overall rates of positive selection, as well as at elucidating the specific genetic changes that underlie adaptations [324]. ***Even if only a small fraction of polymorphisms and substitutions are positively selected, the most interesting biology lies in those exceptions****: for many evolutionary biologists, a sense of awe at the power of Darwinian adaptation is what got us excited about biology in the first place!*

Figure 2.116: **Camouflaged cicada on tree.** *Although the neutral theory provides a powerful framework for modeling molecular evolution,* ***it does not deny the central importance of Darwinian adaptation*** *– in this case driving adaptation of the cicada to be almost perfectly camouflaged in its natural habitat.* Credit: Henk Monster [Link] CC BY 3.0.

Substitution rates and the molecular clock. As shown above, proteins (and DNA sequences) tend to accumulate changes roughly linearly in time, though the rates differ between proteins. This observation would be puzzling if most substitutions are adaptive: why should adaptation occur at a roughly constant rate over hundreds of millions of years, while the organisms themselves, ecosystems, and parameters such as effective population size, vary hugely over time? The Neutral Theory provides a simple model for this.

First, we need to derive the substitution rate for purely neutral sequences. Suppose we sequence a neutral region of the genome in two species that diverged T generations. How many differences do we expect to see be-

tween the two species?

Mutations arise at a rate μ per base pair per generation. Let's look first at fixation events in Species 1. Suppose that the population size of Species 1 is $2N$; then across the entire population of Species 1 we get $2N\mu$ new mutations per base pair each generation. Recall from Chapter 2.1 that new mutations will ultimately fix with a probability equal to their starting frequency: i.e., $1/2N$. Hence, the rate of fixation of mutations is

$$\text{Fixation rate} = [\text{Total rate of new muts}] \times [\text{Fixation prob. of muts}] \quad (2.93)$$

$$= 2N\mu \times \frac{1}{2N} \quad (2.94)$$

$$= \mu \quad (2.95)$$

The population size, N here, cancels out, leading to the crucial result that neutral mutations fix at a rate μ per generation per site, regardless of population size.

Similarly if we compare two species that have diverged for T generations, then at neutral sites the expected frequency of differences is $2\mu T$ [325]. The factor of 2 reflects that fixation events occur in *both* lineages for T generations [326].

Now let's focus specifically on nonsynonymous changes [a]. Think about what happens if a gene contains some positions where mutations would be neutral, and others where mutations would be deleterious: for example mutations in a functional binding pocket of an enzyme might strongly disrupt function, while a change between similarly charged amino acids in an unstructured region might be neutral. Let's suppose that a fraction of λ of all changes are neutral, and $1 - \lambda$ are sufficiently deleterious that they have essentially no chance of fixing [327]. Now we find that mutations fix at a rate

[a] *Recall that nonsynonymous (=missense) substitutions change the amino acid encoded at a position, while synonymous substitutions do not. For example, CCC→GCC changes proline to alanine (nonsynonymous); but CCC→CCG maintains proline (synonymous).*

$$\text{Fixation rate} = \lambda\mu \quad (2.96)$$

per generation, and the expected number of substitutions per site between species is

$$2\lambda\mu T. \quad (2.97)$$

If we convert μ from a per-generation rate to a per-year rate, and assume that this is roughly constant across the phylogeny, and across genes, then *this predicts that substitutions accumulate linearly in time, where the rates are proportional to the fraction of neutral sites. This results in the molecular clock, where the slope is proportional to λ* [328].

$\mathbf{d_n/d_s}$ as an estimator for amino acid constraint. If we want to use Equation 2.97 to estimate λ we need to know both the divergence time T and gene-specific mutation rate μ. Unfortunately we don't always have good estimates of these.

But we can get a better estimator of λ by simply comparing the substitution rates for synonymous and nonsynonymous sites within the same

gene. This is captured in a measure called d_n/d_s (also known as K_a/K_s [329]). Here d_n is the expected number of nonsynonymous substitutions per nonsynonymous site, and d_s is the corresponding number for synonymous substitutions. Then d_n/d_s gives the ratio of the two rates [b].

[b] *Note that d_n and d_s are adjusted for the effective numbers of nonsynonymous and synonymous sites, based on the numbers of possible mutations that would/would not change the encoded protein. Thus d_n/d_s should be 1 in the absence of selection* [330].

To interpret d_n/d_s, let's first make the simplifying assumption that all synonymous mutations are neutral [331]. Then the expected synonymous divergence between two species would $E(d_s) = 2\mu T$, as above.

For nonsynonymous sites in the same gene, the expected nonsynonymous divergence would be $E(d_n) = 2\lambda\mu T$. So the ratio of the expected values tells us that [332]:

$$\frac{d_n}{d_s} = \lambda. \tag{2.98}$$

Notice that since λ represents the *fraction* of neutral sites, d_n/d_s must be between 0 and 1 under this model.

Indeed, this is the case for most genes, as you can see in this plot showing the distribution of d_n/d_s values in mammals [333]:

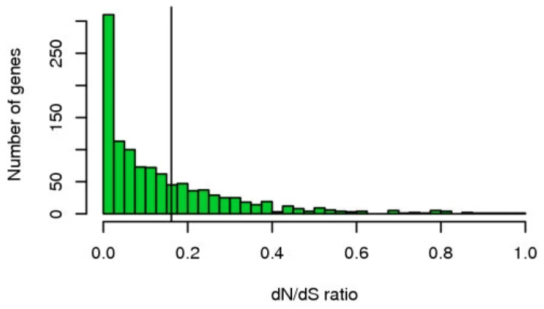

Figure 2.117: **Distribution of $\mathbf{d_n/d_s}$ across human genes.** *The plot shows a histogram of estimated d_n/d_s across genes, measured in the human lineage. The vertical line indicates the mean.* *d_n/d_s is measured on the human lineage since the common ancestor of human, mouse, and pig. Credit: From Fig. 5 of Frank Jørgenson et al (2005) [Link]; CC BY 2.0.*

This study reported a genome-wide average of about $d_n/d_s = 0.14$, which we could interpret to mean that about 14% of amino acid substitutions were effectively neutral.

Positive selection, d_n/d_s, and the MK test. In the absence of positive selection d_n/d_s is always ≤ 1. But what happens if some nonsynonymous mutations are actually favored by selection? Intuitively, you might expect that selection should increase divergence at nonsynonymous sites, and could potentially push $d_n/d_s > 1$. This suggests a test for adaptive evolution of protein sequences: Can we find genes for which d_n/d_s is significantly > 1?

To understand this, let's consider a simple extension of the model to three categories of sites:
- A fraction λ_0 are neutral
- A fraction λ_a are advantageous with selection coefficient s
- A fraction $1 - \lambda_0 - \lambda_a$ are strongly deleterious

Recall that favored mutations fix with probability s [334]. Hence, favored mutations arise at a rate $2N\lambda_a\mu$ per generation, and fix at a rate $2N\lambda_a\mu \cdot s$.

So the expected divergence at nonsynonymous sites in time $2T$ will be $\lambda_0 \cdot 2\mu T + 2N\lambda_a s \cdot 2\mu T$, compared to $2\mu T$ at synonymous sites, and

$$\frac{d_n}{d_s} = \lambda_0 + 2Ns\lambda_a. \tag{2.99}$$

To make this concrete, suppose that $\lambda_0 = 0.2$; and further suppose that 1% of nonsynonymous mutations in a gene have a selective advantage $s = 0.1\%$ in a population of 10^4. Then $2Ns\lambda_a = 0.2$, and d_n/d_s is 0.4. In this example, even though many of the sites are fixed by positive selection, d_n/d_s is still much less than 1.

In fact, we need a *lot* of selection to detect it using d_n/d_s. For example, suppose that 1% of nonsynonymous mutations have an advantage of $s = 1\%$. We now predict that $2Ns\lambda_a = 2.0$ and d_n/d_s will be 2.2, and we reject neutrality.

High $\mathbf{d_n/d_s}$ at MHC genes. It's quite unusual in mammals for selection to be strong enough to drive d_n/d_s above 1, but a famous example occurs in genes of the Major Histocompatibility Complex (MHC) [335]. MHC genes play an **essential role in defense against infection** by presenting fragments of proteins known as peptides for surveillance by T cells. T cells are trained to ignore peptides from our own proteomes; but when they detect foreign peptides they initiate an immune response.

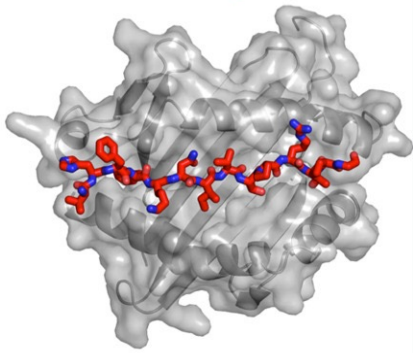

Figure 2.118: **Peptide presentation by MHC.** *MHC proteins (here in gray) play an essential role in the immune system by presenting short peptides (red/blue) for inspection by T cells. MHC proteins must be able to successfully bind a highly diverse and rapidly evolving array of foreign peptides.* *Credit: Figure 3e of Meriem Attaf et al (2015) [Link]. CC BY 4.0*

Crucially, different MHC alleles have different potential binding repertoires. Thus, the universe of peptides that you can present to T cells depends on your genotype across the six MHC genes involved in antigen presentation. There is an overall advantage to having different alleles at each MHC gene, as it expands the potential space of antigens you can present, and particular MHC alleles may especially effective against particular pathogens. All of these factors have led to huge selection pressure for allelic diversity in the MHC, driving ancient balancing selection, similar to the ABO story in the last chapter [336].

Given the strong selection pressure in favor of functional diversity, it should come as no surprise that there is enormous nonsynonymous diversity at functional sites in the MHC genes. A classic 1988 paper by Austin Hughes and Masatoshi Nei examined d_n and d_s between highly diverged human alleles for three MHC genes. They predicted high d_n within the peptide binding region (PBR), but not in the rest of the protein where the function is more conserved [337].

	Peptide Binding (L=57)		Not PBR (L=125)		Exon 4 (L=92)	
	d_n	d_s	d_n	d_s	d_n	d_s
MHC-A	**13.3**	**3.5**	1.6	2.5	1.6	9.5
MHC-B	**18.1**	**7.1**	2.4	6.9	0.5	1.5
MHC-C	**8.8**	**3.8**	4.8	10.5	1.0	2.1

Table 2.7: High $\mathbf{d_n/d_s}$ in MHC genes. *Average d_n and d_s between different human alleles in three MHC genes.* *"Peptide Binding" refers to sites within the PBR; "Not PBR" corresponds to other sites in Exons 2 and 3 that do not contact the peptide; sites in Exon 4 also do not contact the peptide. L indicates numbers of sites. Standard errors for most comparisons were ~2. Modified from Hughes and Nei (1988) [Link].*

Consistent with this logic, you can see above that d_n is larger than d_s within the

peptide binding region, and lower elsewhere. This indicates that there is frequent adaptive evolution within the peptide binding region, and main selective constraint in the structural regions of these genes.

However, more generally, testing for $d_n/d_s > 1$ is not a very powerful test because it's highly unusual to see so many adaptive changes in one small region; secondly, we may not know in advance which sites are likely to be evolving adaptively as we do in the example above (but see [338, 339]).

Tests contrasting polymorphism and divergence. A paper by John McDonald and Martin Kreitman in 1991 suggested a more powerful test for selection by contrasting variation within and between species, now known as the **McDonald-Kreitman or MK test** [340]. The key concept is that selective sweeps occur quickly compared to drift, and so they are more likely to be observed in a data set as differences between species than as polymorphic sites within species.

In one version, the MK test considers gene sequences for multiple individuals from each of two species. Variants can be classified as being either *fixed differences* between the species, or *polymorphic* within one of the species [c]. Similar to the model we used before, the null hypothesis will be that a fraction λ of new nonsynonymous mutations are neutral, and $1 - \lambda$ are strongly deleterious.

[c] *A fixed difference is a variant where all individuals in one species have one variant, while all individuals in the other species have a different variant. A polymorphism would be variable within the sample from one or other species.*

Assuming that selection against deleterious variants is strong enough, we don't expect to see deleterious variants as polymorphisms, and certainly not as fixed differences. In this scenario, we expect the ratio of nonsynonymous to synonymous to be the same (i.e., λ) for both polymorphisms and fixed differences (Panel A):

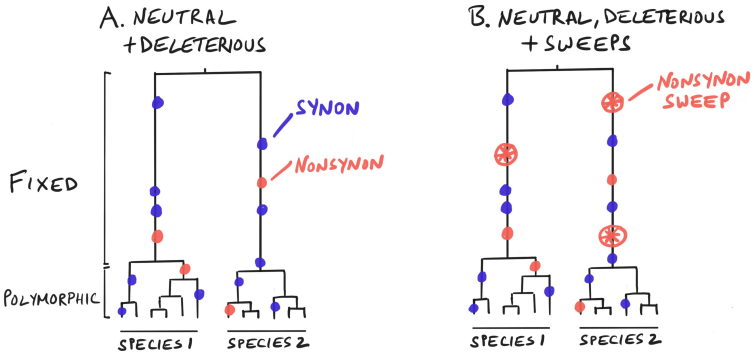

Figure 2.119: **Overview of the MK test. A.** *In the baseline model (Neutral + Strongly deleterious) the expected ratio of nonsynonymous:synonymous variants is the same in the fixed and polymorphic categories.* **B.** *In the model with positive selection, there is a greater fraction of nonsynonymous sites among the fixed differences.*

But if some nonsynonymous sites are positively selected, then these will tend to sweep through populations very quickly (Panel B). Because they sweep quickly, it's rare that one would be just in the process of sweeping right now, and much more likely that they would be fixed differences [d]. For this reason we expect that *positively selected variants will increase the fraction of nonsynonymous variants among the fixed differences.*

[d] *You can look at Figure 2.95 to see that selected variants fix much faster than neutral variants.*

Consistent with the positive selection model, the first application of the MK approach found a much higher fraction of nonsynonymous substitutions *between* Drosophila species (41% of substitutions) compared to nonsynonymous polymorphisms within species (just 5% of variants):

	Fixed	Polymorphic
Nonsynonymous	7	2
Synonymous	17	42
% Nonsynon.	**41.2%**	**4.7%**

Table 2.8: Excess of nonsynonymous substitutions in the Drosophila ADH gene. *The table shows the numbers of nonsynonymous and synonymous variants that are either polymorphic within species, or fixed between species (P-value for a test of independence is* 0.006*).* *Modified from McDonald and Kreitman (1991) [Link].*

The authors interpreted this as evidence that positive selection at ADH drives nonsynonymous fixations that accumulate as an excess of between-species differences [341] [342].

Since then, MK analyses in the genome-wide era have revealed rampant positive selection in Drosophila: likely as many as 50% of nonsynonymous differences between species were fixed by positive selection [343].

For humans, in contrast, it seems that a much smaller fraction of nonsynonymous differences between humans and other primates were fixed by positive selection: likely in the range of 0–10%, although the precise fraction is still a matter of debate [344]. *This work shows that the great majority of nonsynonymous substitutions in primates are effectively neutral.*

Linked selection: background selection and hitchhiking. This brings us to our last major selection topic, **linked selection**, which deals with the effects of selection–both positive and negative–at nearby sites. Here, we ask: *How does selection affect the patterns of genetic diversity at nearby neutral sites* [345]*?*

Our story begins around the same time as development of the McDonald-Kreitman test, when an observation from Drosophila presented an important new challenge to the Neutral Theory. A 1992 paper by David Begun and Chip Aquadro showed that regions of the fruit fly (Drosophila) genome with low recombination rates tend to have low genetic diversity [346].

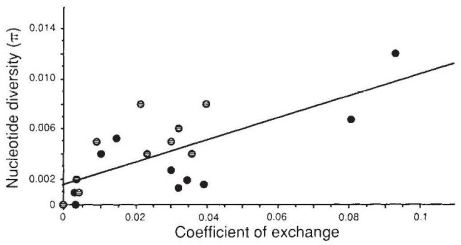

Figure 2.120: **Classic plot of the relationship between recombination rate and genetic diversity in fruit flies (1992).** *The x-axis shows a measure of local recombination rate; the y-axis is average pairwise heterozygosity, π; each data point is a different sequenced locus. The null hypothesis that the slope is* 0 *is rejected with* $p = 0.0007$. *Credit: Fig. 1 of David Begun and Charles Aquadro (1992) [Link] Used with permission.*

Similar patterns are also seen in humans (side panel) [347].

Begun and Aquadro proposed that this observation is evidence for widespread **genetic hitchhiking with selective sweeps**. Recall from Chapter 2.6 that when a favored mutation sweeps rapidly through the population, it carries a surrounding haplotype with it, up to high frequency, in a process known as hitchhiking. As a sweep nears completion it eliminates

genetic variation in a window around the selected site (Figure 2.96).

The size of the window is inversely related to $r \times 2log(2N)/s$, where r is the local recombination rate, N is the population size, and s is the selection coefficient. This is intuitive: **when r is high, recombination breaks up the sweeping haplotype much more efficiently than when r is low.**

Begun and Aquadro hypothesized that sweeps are scattered randomly across the genome. When they sequenced a locus with low recombination rate it was much more likely to fall within the window of a recent sweep (and therefore, have low diversity) than when they sequenced a locus with high recombination rate. They concluded that "Hitch-hiking thus seems to occur over a large fraction of the *Drosophila* genome and may constitute a major constraint on levels of genetic variation".

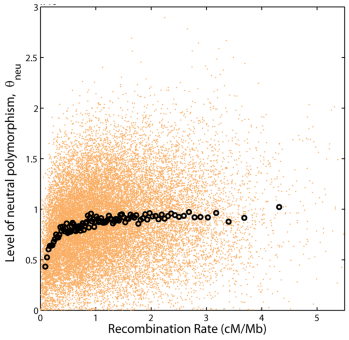

Figure 2.121: **Human genetic diversity is also reduced in regions of low recombination.** *Black dots show binned averages of the genetic diversity ($\theta \times 10^{-3}$) as a function of local recombination rate. Orange dots show raw data in a sliding window across the genome.* Credit: *Fig. 1B of James Cai et al (2009); CC BY 4.*

Background selection. However, the next year an alternative explanation, dubbed *background selection*, was proposed by Brian Charlesworth and colleagues [348]. The essential concept of background selection is that when deleterious mutations arise, they may drift briefly but are unlikely to contribute to the population long-term. As those variants are eventually purged, any linked neutral variants are lost too.

One helpful way to think about this is that, at any given locus, *the copies of this locus present in the population today are primarily descended from past copies of the locus that did not carry deleterious mutations. Thus, deleterious mutations can be thought of as reducing effective population size within a linked region.*

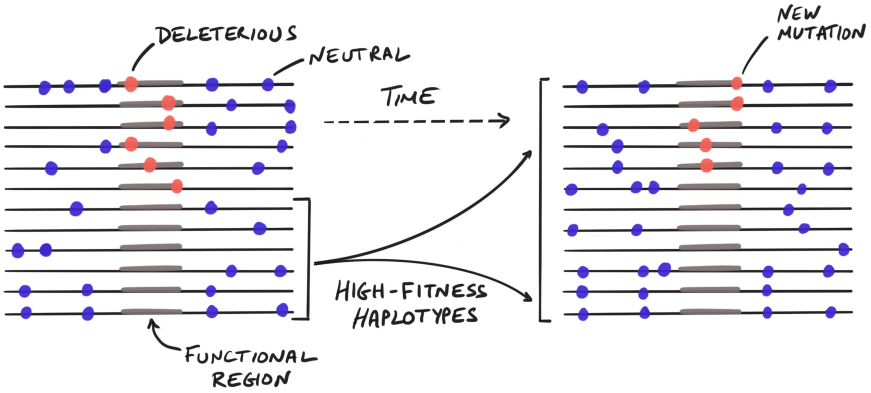

Figure 2.122: **Background selection.** *At any given time a fraction of chromosomes carry deleterious variants (red mutations). These chromosomes have low fitness and don't contribute much to future populations in the long term. Neutral variants linked to deleterious variants will eventually be removed by selection. Meanwhile new deleterious variants continue to arise by mutation.*

To model this, let's first look at background selection in a region without recombination [349]. Define f as the total fraction of chromosomes that carry deleterious mutations. A simple model [350] suggests that at equilibrium

$$f \approx L\mu/hs, \tag{2.100}$$

where L is the number of basepairs that can produce deleterious mutations, μ is the mutation rate per base pair, and hs is the selective disadvantage for a heterozygote [e].

Provided that selection is strong enough that individual deleterious mutations don't persist long in the population ($hs >> 1$), you can think of

[e] *In a minor abuse of notation, in this section we use $hs > 0$ to indicate a selective disadvantage. The derivation requires that selection is considerably stronger than drift, i.e., $hs >> 1/N$.*

this as reducing the effective population size locally by a factor $1 - f$. Then we can approximate the expected pairwise genetic diversity, $\mathrm{E}[\pi]$, in this region as

$$\mathrm{E}[\pi] = \pi_0(1 - \frac{L\mu}{hs}), \tag{2.101}$$

where $\mathrm{E}[\pi_0]$ is what the expected genetic diversity would have been if there were no background selection.

What happens with recombination? Let's say we sequence a region that is at a recombination fraction r from a conserved functional element. Now things are more complicated, because a neutral variant in the sequenced region could be rescued by recombining away from a linked deleterious mutation. After a flurry of math [351], the expected diversity is found to be

$$\mathrm{E}[\pi] = \pi_0\left[1 - \frac{L\mu}{hs(1 + r/hs)^2}\right]. \tag{2.102}$$

The last part of this expression, $\frac{L\mu}{hs(1+r/hs)^2}$, represents the proportional decrease in variation due to background selection. *This has the intuitive form that the impact of background selection increases with the deleterious mutation rate $L\mu$, and decreases with recombination distance r.* The relationship with selection strength is more complicated [352].

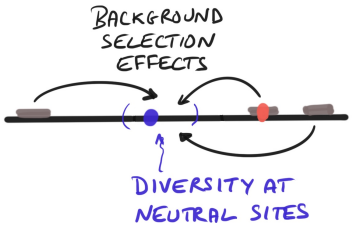

Figure 2.123: **Genetic diversity at neutral sites is reduced by background selection from linked functional elements.** *Each functional element reduces expected diversity by a factor of* $(1 - \frac{L_i\mu}{hs(1+r_i/hs)^2})$.

Next, the total strength of background selection experienced at a site depends on the cumulative contributions from all linked functional loci (for example, all coding exons, conserved gene regulatory elements, etc). The total reduction in π is a product of the contributions from each functional element:

$$\mathrm{E}[\pi] \approx \pi_0 \prod_{i=1}^{M}\left[1 - \frac{L_i\mu}{hs(1 + r_i/hs)^2}\right]. \tag{2.103}$$

where i indexes each of M linked functional elements [353].

Does this model fit real data?

In a 2009 paper, Graham McVicker and colleagues used this approach to predict the background selection effect of constrained regions across the human genome [354]. As you see from Equation 2.103, the strength of background selection at any specific location depends on the local landscape of linked functional elements. This can be used to predict genetic diversity at neutral sites across the genome, depending on the number, size, and genetic distance to nearby functional elements in the genome sequence. This reduction in diversity is commonly written as **B** [355]:

$$B = \frac{\mathrm{E}[\pi]}{\pi_0} \tag{2.104}$$

The plot below shows an updated version of McVicker's analysis [356]. As you can see, the background selection model provides a remarkably good prediction of the landscape of genetic diversity in humans:

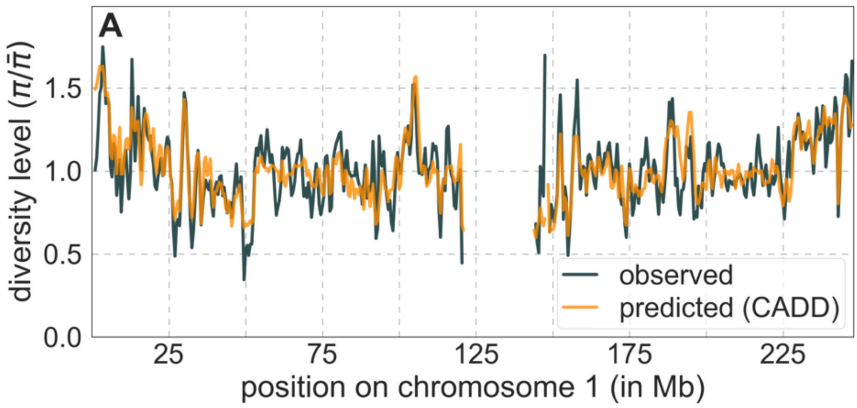

Figure 2.124: **Human genetic diversity predicted from background selection model.** *Genetic diversity along chromosome 1 is plotted in teal; predictions from a background selection model are in orange.* *Data are for Yoruba (from Nigeria). The y-axis is genetic diversity π divided by the genome-wide average. The data are plotted in megabase-sized windows. The gap at the center of the plot is due to repetitive regions near the centromere. Credit: Figure 2 from David Murphy et al (2021) [Link]. CC BY 4.*

This is actually quite a dramatic effect, with the model accounting for most variation in genetic diversity (at megabase scale) across the genome, emphasizing the importance of linked selection in shaping genetic variation.

Background selection or hitchhiking? Thus, in both humans and flies (and other species) we see a strong relationship between neutral sequence diversity, and the local recombination rate and density of nearby functional sequence. This is compelling evidence that linked selection plays a major role in shaping genetic diversity across the genome. But it leaves us wondering how much of the linked selection effect is due to background selection versus hard sweeps [357].

One way of distinguishing these is to note that if hard sweeps are important, then there should be a dip in diversity specifically near the sites of completed sweeps. This is different from the general depletion of variation due to background selection. We don't know which sites have had recently completed sweeps, but we could hypothesize that these would be enriched at recent nonsynonymous substitutions. Under this hypothesis, diversity near nonsynonymous substitutions would reflect a mixture of neutral and selected signals. To summarize: *if an appreciable fraction of nonsynonymous substitutions on the human lineage are recently completed hard sweeps, then we should see lower diversity near those sites compared to a model with background selection only* [358].

But, instead, the genetic diversity around nonsynonymous substitutions can be predicted entirely from the background selection model. The plot below shows the average of genetic diversity around all nonsynonymous substitutions, along with the predictions under background selection. There's a dip at the center of the plot, but this is only because nonsynonymous sites are generally in regions with lots of functional sequences – as you can see the data are extremely similar to the background selection model:

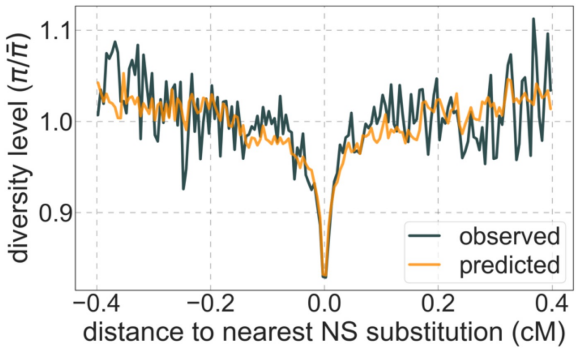

Figure 2.125: **Average levels of genetic diversity at nonsynonymous substitutions and comparison to predictions from background selection.** *Genetic diversity at nonsynonymous substitutions (teal) is accurately predicted from a background selection model (orange). The close fit at nonsynonymous sites argues against frequent hard sweeps driving nonsynonymous substitutions.* *Credit: Figure 3 from David Murphy et al (2021) [Link]. CC BY 4.*

This analysis would have had power to detect a signal if as much as 10% of nonsynonymous variants had swept with strong selection (s=1%) [359], though it has less power to detect soft sweeps. In contrast, this type of analysis *does* show a clear signal in Drosophila, where selective sweeps seem to be much more common [360].

In summary, this analysis and the MK results argue that at most a small fraction (<10%) of nonsynonymous variants in humans were fixed by strong positive selection. Similarly, genomewide selection scans reveal relatively few unequivocal examples of recent sweeps in noncoding regions. This leaves open the possibility of positive selection through soft sweeps, much weaker hard sweeps, and polygenic adaptation as we'll discuss next.

Concluding remarks. In this section of the book we have covered some of the core principles for understanding genetic variation. One remarkable aspect of population genetics is that many of the fundamental concepts extend logically from the basic genetic and population processes: mutation; Mendelian segregation; linkage; random mating and population structure; and different forms of selection.

That said, while many key concepts were already understood 50 years ago, it has taken much longer to determine the relative importance of the different processes–in particular the impact of genetic drift, linkage, and the different types of selection in shaping patterns of variation and evolution–and many aspects of this are important areas of research now that we have much richer genome data, and modern tools from functional genomics.

I think it's fair to say that versions of the Neutral Theory now provide the central structure for most models of genetic variation: at least 90% of new single nucleotide mutations are essentially neutral, and most of what is not neutral is deleterious. However, we also now know that linked selection in the genome, mainly from background selection, is pervasive, so that diversity in most of the genome is reduced relative to the maximum possible under a fully neutral model.

What then, is the role of positive selection? Even if only a tiny fraction of variants are positively selected, we do know that the natural world, including humans, show an astonishing diversity of forms. Organisms are

amazing molecular machines, and exquisitely adapted to their environments. This must happen through forms of adaptation. As I discussed at length, we do now have compelling examples of the various forms of positive selection acting in humans: including hard and soft sweeps, and ancient balancing selection.

However, my personal reading of the data is that strong hard selection on individual loci has been rare in the human genome during the past $\sim$ 200,000 years when we can best detect it. Many of the exceptions where we do see sweep signals are at genes where a single protein plays an exceptional role in some process–for example Duffy, which serves as a specific receptor for vivax malaria; or lactase which plays an essential role in digesting lactose. The relative importance of different modes of selection seems to vary greatly across species, and hard sweeps may be less important in humans than in some other species that have been studied, including flies and stickleback fish.

It's possible that environmental pressures acting on human populations are often variable and inconsistent across space and time, and thus it is rare for selection be both strong and consistent enough over the many thousands of years that are required for hard sweeps in a species with our long generation time. This hypothesis may be consistent with recent work on ancient DNA identifying many short-term selective frequency shifts [361]. This work suggests that perhaps much of the recent selection has taken the form of partial soft sweeps – which would not show up clearly in most analyses [362].

Lastly, it is likely that most human adaptation comes through polygenic shifts of complex traits. We do know that the genetic variation in most phenotypes, aside from monogenic genetic diseases, is highly polygenic. It must surely be the case that environmental pressures are continually pushing optimal phenotypes around in some high-dimensional phenotype space as conditions change. However, polygenic adaptation leaves little trace in the data and, at the time of writing, detection remains difficult [363]. We will consider the population genetics of polygenic traits in detail later in the book.

Well done! You have now completed the population genetics section of the book! These main principles are useful for understanding all aspects of human genetic variation. In the next section we'll focus on application of these principles for understanding the genetic structure and history of human populations.

Figure 2.126: **Exquisite adaptation of the spicebush swallowtail caterpillar.** *This caterpillar discourages would-be predators using pigmented spots that mimic snake eyes.* *Credit: Michael Hodge [Link] CC BY 2.*

Notes and References.

[310]A short history of population genetics:
Charlesworth B, Charlesworth D. Population genetics from 1966 to 2016. Heredity. 2017;118(1):2-9

[311]In 1963 Dick Lewontin who, a few years later, helped introduce electrophoresis into population genetics, lamented the plight of population genetics in the absence of data: *"In many ways the lot of the theoretical population geneticist of 1963 is a most unhappy one. For he is employed, and has been employed for the last thirty years, in polishing with finer and finer grades of jeweler's rouge these three colossal monuments of mathematical biology...By the end of 1932 Haldane, Fisher, and Wright had said everything of truly fundamental importance about the theory of genetic change in populations and it is due mainly to man's infinite capacity to make more and more out of less and less, that the rest of us are not currently among the unemployed."* As quoted in Singh and Krimbas, Evolutionary Genetics: From molecules to morphology, Chapter 11; the original does not seem to be online.

[312]A short history of electrophoresis:
Charlesworth B, Charlesworth D, Coyne JA, Langley CH. Hubby and Lewontin on protein variation in natural populations: when molecular genetics came to the rescue of population genetics. Genetics. 2016;203(4):1497-503

[313]Harris H. C. Genetics of Man Enzyme polymorphisms in man. Proceedings of the Royal Society of London Series B Biological Sciences. 1966;164(995):298-310
Hubby JL, Lewontin RC. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. Genetics. 1966;54(2):577
Lewontin RC, Hubby JL. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics. 1966;54(2):595

[314]Charlesworth et al (2016).

[315]One viewpoint, motivated by observations of balanced inversion polymorphisms in Drosophila pseudoobscura, by Dobzhansky, emphasized the importance of balancing selection.

[316]Lewontin and Hubby (1966).

[317]Zuckerkandl and Pauling called this the "molecular evolutionary clock", though this is usually shortened to "molecular clock" in modern usage [REF]. See also the Kumar NRG review 2005:
Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. Journal of theoretical biology. 1965;8(2):357-66
Kumar S. Molecular clocks: four decades of evolution. Nature Reviews Genetics. 2005;6(8):654-62

[318]Dickerson RE. The structure of cytochrome c and the rates of molecular evolution. Journal of Molecular Evolution. 1971;1:26-45

[319]King JL, Jukes TH. Non-Darwinian Evolution: Most evolutionary change in proteins may be due to neutral mutations and genetic drift. Science. 1969;164(3881):788-98.

[320]Two key papers in 1968 helped to outline this: Kimura (1968); King and Jukes (1968). In the longer run, Kimura became most influential due to his continued work on this, including his 1983 book.
Kimura M, et al. Evolutionary rate at the molecular level. Nature. 1968;217(5129):624-6

[321]The quotes are from the Introduction to Kimura (1983):
Kimura M. The neutral theory of molecular evolution. Cambridge University Press; 1983

[322]One recent review is strongly critical of the Neutral Theory, in part for under-appreciating the role of linked selection:
Kern AD, Hahn MW. The neutral theory in light of natural selection. Molecular biology and evolution. 2018;35(6):1366-71
however, to the extent that the linked selection signal is due to background selection it can actually be viewed as a natural extension of the Neutral Theory:
Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, et al. The importance of the neutral theory in 1968 and 50 years on: a response to Kern and Hahn 2018. Evolution. 2019;73(1):111-4

[323]So far we have been following the Neutral Theory in treating mutations as either neutral, or strongly deleterious. However, starting in 1973, another Japanese scientist Tomoko Ohta emphasized the role of nearly-neutral mutations in protein evolution (Ohta 1973 paper, and later Annals review). In contrast to this simplest model, she argued that many amino acid substitutions may be weakly selected – i.e., with $|2Ns|$ around 1 or less. Notice that the "drift barrier" model discussed earlier is closely related to this model. The Nearly Neutral model allows for much more complexity in protein evolution: for example we can expect higher substitution rates in populations with smaller effective population sizes. Hence in the Nearly Neutral model, λ, the fraction of approximately neutral sites, is no longer a fixed property of a gene,

but instead increases or decreases depending on changes in N_e. Secondly, the fixation of nearly neutral mutations can lead to clumping of substitutions over time, because the substitution of one weakly deleterious mutation may be followed by substitution of weakly advantageous compensatory mutations nearby.

[324]Sackton TB. Studying natural selection in the era of ubiquitous genomes. Trends in Genetics. 2020;36(10):792-803

[325]Technically, here, T is the average coalescent time for lineages from each of the two species, rather than the species split time.

[326]Note that in data analysis, the number of sequence differences between two species is actually a lower bound on the number of substitutions, as there may be "multiple hits": i.e., positions that have had multiple substitutions; there are many statistical methods to account for this.

[327]Variants are sufficiently deleterious that they have essentially no chance of fixing if $s << -1/N$.

[328]It's long been observed that the molecular clock is not *precisely* clocklike. The strongest version of the molecular clock model would suggest that substitutions occur at a constant rate, uniformly in time (technically, as a Poisson Process with a fixed rate). In practice, substitutions tend to be more clumped within a phylogeny than expected under the ideal clock model; this is referred to as the **overdispersed molecular clock**. Early work documenting this argued that the overdispersed clock was evidence against the Neutral Model, and in favor of bursts of adaptive evolution Gillespie (1989) but later work has argued that much of this can be explained by a combination of effects, including gene- and lineage-specific changes in mutation rates, as well as substitutions of nearly neutral mutations, as in Ohta's Nearly Neutral Theory. For recent work in this area see work from Bedford and colleagues. Note that Bedford et al found stronger overdispersion at nonsynonymous sites than synonymous, indicating that these are not purely mutational effects. Secondly they found stronger overdispersion in mammals than in flies, than in yeast; this pattern suggests that overdispersion may be stronger in small populations than in large populations, which is perhaps the opposite of what we might expect if the overdispersion were mainly due to bursts of adaptation.

Gillespie JH. Lineage effects and the index of dispersion of molecular evolution. Molecular biology and evolution. 1989;6(6):636-47

Bedford T, Wapinski I, Hartl DL. Overdispersion of the molecular clock varies between yeast, Drosophila and mammals. Genetics. 2008;179(2):977-84

Bedford T, Hartl DL. Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in Drosophila. Molecular biology and evolution. 2008;25(8):1631-8

[329]The traditional notation dN/dS or d_N/d_S notation introduces multiple notational clashes: d is a distance and not a derivative; N and S refer to nonsynonymous and synonymous sites and not population size or selection. For this reason I use lower case, subscript n and s. In general the usage should hopefully be clear from context.

[330]Here I'm skating over many complexities in estimating d_n/d_s. First, it varies across papers whether these distances are treated as expected outcomes of an evolutionary process, or the realized numbers of substitutions. Even if it's the latter, these are still difficult to estimate due to the possibilities of multiple substitutions occurring at the same sites, and variation in the rates of transitions, transversions, and other mutation types. Lastly, one should be cautious when estimating ratios of random variables – for example the simple estimator can blow up for short genes if we don't observe any synonymous substitutions.

[331]You might reasonably worry about non-neutral effects on synonymous sites, including codon bias, or exonic splicing enhancers that overlap synonymous sites; but in aggregate these are generally weak compared to selective constraint on amino acid sequences so using synonymous sites as a baseline is generally a useful approximation.

[332]In practice d_n/d_s is usually estimated as a ratio of estimates, namely $\hat{d_n}/\hat{d_s}$. Interpreting this is a bit more tricky because obviously the estimate comes with sampling variation, and as a ratio of random variables the estimator is a biased estimator of λ.

[333]Jørgensen FG, Hobolth A, Hornshøj H, Bendixen C, Fredholm M, Schierup MH. Comparative analysis of protein coding sequences from human, mouse and the domesticated pig. BMC biology. 2005;3:1-15

[334]Chapter 2.5. As before we take $h = 0.5$

[335]In humans the MHC is also known as the HLA or Human Leukocyte Antigen complex. The MHC/HLA is the main focus for transplant matching in organ donations because it is essential for distinguishing self from non-self antigens. The MHC is also the major driver of autoimmune disease – the immune system treads a delicate balance between sensitive immune surveillance for pathogens versus the risk of autoimmunity.

[336]Like at ABO, distinct allelic lineages have likely persisted for $>$ 20 million years, and there is enormous genetic diversity in the MHC region, with nucleotide diversity reaching well above 1% – more than 10-fold the genome-wide average.background;

Jensen JM, Villesen P, Friborg RM, Mailund T, Besenbacher S, Schierup MH, et al. Assembly and analysis of 100 full MHC haplotypes from the Danish population. Genome research. 2017;27(9):1597-607

Norman PJ, Norberg SJ, Guethlein LA, Nemat-Gorgani N, Royce T, Wroblewski EE, et al. Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class I and II. Genome research. 2017;27(5):813-23

[337]Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 1988;335(6186):167-70.

[338]One other fascinating example of high d_n/d_s is found in the PRDM9 zinc fingers, which you will recall from Chapter 2.3 play the critical role of directing recombination events:

Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, et al. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS genetics. 2009;5(12):e1000753.

[339]For this reason there has been a great deal of work on improving power to identify particular sites that are subject to positive selection, even within genes that are constrained at most positions eg:

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution. 2007;24(8):1586-91

Sackton TB. Studying natural selection in the era of ubiquitous genomes. Trends in Genetics. 2020;36(10):792-803.

[340]McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351(6328):652-4.

The MK test built on other contemporaneous work, including notably the HKA test

Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. Genetics. 1987;116(1):153-9

[341]It's beyond our scope here, but there has been a great deal of work on more complicated models that extend this basic idea. One weakness of the original MK test is that it ignores the fact that deleterious variants are much more likely to be polymorphic than to be substitutions: this in turn reduces power to detect an excess of nonsynonymous substitutions. However, it's possible to improve the test by considering only common variants, or to use the polymorphism data to estimate a distribution of selection coefficients to make more-powerful MK tests, eg:

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, Glanowski S, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005;437(7062):1153-7

Messer PW, Petrov DA. Frequent adaptation and the McDonald–Kreitman test. Proceedings of the National Academy of Sciences. 2013;110(21):8615-20

[342]However it's worth noting that as the tests become more powerful, they also become more sensitive to model assumptions. One key vulnerability is variation in ancestral population sizes: for example, a small ancestral population size could allow more weakly deleterious variants to fix, and conversely for a large ancestral population size:

Eyre-Walker A. Changing effective population size and the McDonald-Kreitman test. Genetics. 2002;162(4):2017-24

[343]Sella G, Petrov DA, Przeworski M, Andolfatto P. Pervasive natural selection in the Drosophila genome? PLoS genetics. 2009;5(6):e1000495

Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular biology and evolution. 2009;26(9):2097-108

[344]Eyre-Walker and Keightley (2009) write that analysis of the human data "...reveals little evidence for adaptive substitutions. However, the true frequency of adaptive substitutions in human-coding DNA could be as high as 40%, because estimates based on current polymorphism may be strongly downwardly biased by a decrease in the effective population size along the human lineage." Boyko et al (2008) estimated 9% in their baseline model. Uricchio et al (2019) estimated 13%. Again, it's important to take all of these estimates with caution as the MK test is easily misled by changes in N_e, which affect the rates of fixation of nearly neutral variants.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS genetics. 2008;4(5):e1000083

Uricchio LH, Petrov DA, Enard D. Exploiting selection at linked sites to infer the rate and strength of adaptation. Nature ecology & evolution. 2019;3(6):977-84

[345]Interactions between selected sites, or between selected sites and nearby neutral sites are sometimes referred to as **Hill-Robertson interference**, based on early work showing that selection at linked sites tends to reduce the efficacy of selection at both sites.

Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genetics Research. 1966;8(3):269-94

Felsenstein J. The evolutionary advantage of recombination. Genetics. 1974;78(2):737-56

[346]Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature. 1992;356(6369):519-20

[347]Cai JJ, Macpherson JM, Sella G, Petrov DA. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS genetics. 2009;5(1):e1000336

[348]Charlesworth B, Morgan M, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289-303

Charlesworth B. Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture. Journal of Heredity. 2013;104(2):161-71

[349]Theory on background selection: Charlesworth et al (1993);

Hudson RR, Kaplan NL. Deleterious background selection with recombination. Genetics. 1995;141(4):1605-17

Nordborg M, Charlesworth B, Charlesworth D. The effect of recombination on background selection. Genetics Research. 1996;67(2):159-74

Elyashiv E, Sattath S, Hu TT, Strutsovsky A, McVicker G, Andolfatto P, et al. A genomic map of the effects of linked selection in Drosophila. PLoS genetics. 2016;12(8):e1006130

Buffalo V, Kern AD. A Quantitative Genetic Model of Background Selection in Humans. bioRxiv. 2023:2023-09

[350]Note that for consistency with the background selection literature, and to simplify the notation, we use $s > 0$ in this section to indicate a deleterious allele, i.e., that fitnesses $1, 1-hs, 1-s$, with $h \in (0,1]$ and $s > 0$ indicate a deleterious derived allele.

We can solve for f by noting that the input of new deleterious mutations per generation is $2NL\mu$, and the number of deleterious mutations removed by selection is $N \cdot 2f(1-f) \cdot hs$ (the latter uses Hardy Weinberg, assuming that f is low enough that most deleterious mutations are heterozygous). At equilibrium, input equals output, and solving for f we get $f \approx 2NL\mu/hs$.

[351]This is Equation 11 from Nordborg et al (1996); see also Hudson and Kaplan (1994)

[352]This expression implies the interesting result that for a fixed distance r, the background selection effect is strongest when $hs = r$. In other words, at nearby functional elements (small r), small values of hs remove the most variation because the deleterious variants can drift up to become relatively common before ultimately being removed. But at large distances, only strong selection really matters: if selection is weak the linked variants have time to recombine to other chromosomes. Thus, assuming a recombination rate of 1cM/MB, at 100kb from a function region, weakly deleterious variants with $hs = 0.1\%$ would have the most impact but at 1MB distance variants with stronger effects, $hs = 1\%$, would have the most impact.

[353]Coop 2020 Eq 13.13; Nordborg, Elyashiv et al (2016) Eq 2. Note that for computational purposes it is common to use the further approximation that $1 - x \approx e^{-x}$ and then to rewrite this in the form $\exp\sum x_i$.

[354]McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS genetics. 2009;5(5):e1000471

[355]It's sometimes known as McVicker's B, which is an example of Stigler's Law of Eponymy [Link].

[356]Murphy DA, Elyashiv E, Amster G, Sella G. Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. Elife. 2022;12:e76065

[357]For examples of contrasting views see Lohmueller et al (2011), Enard et al (2014)

Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, et al. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS genetics. 2011;7(10):e1002326

Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome research. 2014;24(6):885-95

and additional references as follow.

[358]This method was pioneered by

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. science. 2011;331(6019):920-4

Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in Drosophila simulans. PLoS genetics. 2011;7(2):e1001302

Here I present results from the updated analysis by Murphy et al (2021).

[359]Or 25% with moderate selection (s=0.1%). The power analyses are from Hernandez (2011)

[360]Elyashiv et al (2016) estimated that 4% of missense substitutions were fixed by strong selection, and 35% by weak selection.

[361] Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528(7583):499-503

[362] There is a large literature on selection scans in humans and primates, using a variety of analysis techniques and data, and reaching different conclusions on the frequency, strength, and types of selection. Some of these discrepancies may reflect poor calibration of some studies, but my suspicion is that much of this probably reflects a lot of weak, soft, selection that forces variants up or down in frequency but rarely to fixation. This would lead to low power and poor replication across study types. It's plausible that a lot of this selection is actually the tail-end of the distribution of polygenic effects.

[363] Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced signal for polygenic adaptation of height in UK Biobank. Elife. 2019;8:e39725

Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife. 2019;8:e39702

Cox SL, Ruff CB, Maier RM, Mathieson I. Genetic contributions to variation in human stature in prehistoric Europe. Proceedings of the National Academy of Sciences. 2019;116(43):21484-92

Chen M, Sidore C, Akiyama M, Ishigaki K, Kamatani Y, Schlessinger D, et al. Evidence of polygenic adaptation in Sardinia at height-associated loci ascertained from the Biobank Japan. The American Journal of Human Genetics. 2020;107(1):60-71

Hayward LK, Sella G. Polygenic adaptation after a sudden change in environment. Elife. 2022;11:e66697