Supplementary Materials

Vince Buffalo and Andrew Kern

June 25, 2023

Contents

1 Theory

1.1 Overview of Quantitative Genetic Models of Effective Population Size

Here, we step through a quick derivation of **Santiago1995-hx**. Throughout, we assume random mating, hermaphroditic individuals, and a constant population size. The change in a neutral allele's frequency in one generation can be partitioned into the three sources of stochasticity: the random associations with fitness backgrounds (i.e. *draft*), the non-heritable randomness in family size, and the Mendelian noise from heterozygotes segregating. If we let $x_{0,i} \in \{0, 1/2, 1\}$ be the frequency of neutral alleles individual i in generation 0 carries, we can partition the random neutral allele frequency of the population (x_t without the individual index) into the underlying stochastic causes,

$$x_1 = \frac{1}{2N} \sum_{i=1}^{N} \left(x_{0,i} k_{0,i} + \sum_{j=1}^{k_{0,i}} \delta_{0,i,j} \right) \tag{1}$$

where $k_{0,i}$ is the number of surviving gametes parent i passes on, and $\delta_{0,i,j}$ is a random term that encapsulates the noise due to Mendelian segregation of heterozygotes. If parent i is a homozygote ($x_{0,i} \in \{0, 1\}$), then $\delta_{0,i,j} = 0$, whereas if $x_{0,i} = 1/2$ then $\delta_{0,i,j} = \pm 1/2$ with equal probability. This is because a heterozygous parent will transmit half a neutral allele in expectation, but each round of Mendelian segregation must pass on either 0 or 1 alleles, which the random $\pm 1/2$ term imposes. The factor of $1/2$ is due to the fact that we're summing over N diploids, but considering the number of gametes they transmit. Since each diploid parent must have two offspring to maintain a constant population size, $1/N \sum_i k_{0,i} = 2$.

The frequency in the initial generation is $x_0 = 1/N \sum_{i=1}^{N} x_{0,i}$, though to indicate that we treat this as fixed, we use $p_0 = x_0$. Then, the allele frequency change is,

$$\Delta x_1 = x_1 - p_0 = \frac{1}{2N} \sum_{i=1}^{N} x_{0,i}(k_{0,i} - 1) + \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{k_{0,i}} \delta_{0,i,j} \tag{2}$$

$$\Delta x_1 = K_1 + H_1 \tag{3}$$

where K_1 and H_1 are the random change in neutral allele frequency change due to offspring number (including heritable and non-heritable components), and Mendelian segregation in generation 1.

Now, let us look the variance of $\mathrm{var}(\Delta x_1)$ over evolutionary replicates. Since the Mendelian segregation and the offspring random process are independent,

$$\mathrm{var}(\Delta x_1) = \mathrm{var}(K_1) + \mathrm{var}(H_1). \tag{4}$$

Looking at each term,

$$\mathrm{var}(K_1) \approx \frac{1}{4N^2} \sum_{i=1}^{N} \mathrm{var}\left(x_{0,i}(k_{0,i} - 1)\right) \tag{5}$$

where we ignore the covariance terms due to the sum, since these are on order $1/N^3$. In the first generation from an arbitrary starting point, the neutral alleles assort independently into diploids with respect to their fitness, so we can simply the variance as

$$\mathrm{var}(K_1) \approx \frac{1}{4N^2} \sum_{i=1}^{N} \mathrm{var}(x_{0,i})\,\mathrm{var}(k_{0,i} - 1) \tag{6}$$

$$\approx \frac{1}{4N}\,\mathrm{var}(x_{0,i})\,\mathrm{var}(k_{0,i}). \tag{7}$$

Assuming no correlation between parental gametes (e.g. no inbreeding), $\mathrm{var}(x_{0,i})$ is the binomial variance in individual allele frequency, or $p_0(1-p_0)/2$, and $\mathrm{var}(k) := \mathrm{var}(k_{0,i})$ is the offspring variance of an individual given by the reproduction process. For example, if the reproduction process is a neutral multinomial Wright–Fisher, $\mathrm{var}(k) \approx 2$. Then, the variance in allele frequency change due to non-heritable offspring number variation is,

$$\mathrm{var}(K_1) \approx \frac{p_0(1-p_0)}{2N}\frac{\mathrm{var}(k)}{4}. \tag{8}$$

The Mendelian noise variance term can be derived similarly. First, the sum over each parent's transmitted gametes can be simplified by noting that parents are exchangeable over evolutionary replicates with respect to their contribution to this term. In a constant population size, the double summation over parents and their offspring can be replaced by a summation over offspring, since both sum N exchangeable terms. Then, note that $\mathrm{var}(\delta_{i,j}) = 1/4 p_0(1-p_0)$, so

$$\mathrm{var}(H_1) = \frac{p_0(1-p_0)}{2N}\frac{1}{2}. \tag{9}$$

Finally, we have

$$\mathrm{var}(\Delta x_1) = \mathrm{var}(K_1) + \mathrm{var}(H_1) \tag{10}$$

$$\approx \frac{p_0(1-p_0)}{2N}\frac{\mathrm{var}(k)}{4} + \frac{p_0(1-p_0)}{2N}\frac{1}{2} \tag{11}$$

$$\approx \frac{p_0(1-p_0)}{2N}\left(\frac{\mathrm{var}(k)}{4} + \frac{1}{2}\right) \tag{12}$$

(c.f. **Santiago1995-hx** equation 2 and **Buffalo2019-qs** equation 30). Note that if the reproduction process is a neutral multinomial Wright–Fisher, $\mathrm{var}(k) \approx 2$, and this simplifies to the expected Wright–Fisher variance. If we were to *define* a variance effective population size by setting this variance in neutral allele frequency change to the expected variance under a Wright–Fisher model (V_{WF}), we'd have

$$V_{\mathrm{WF}} := \frac{p_0(1-p_0)}{2N_e} \tag{13}$$

$$V_{\mathrm{WF}} = \mathrm{var}(\Delta p_1) \tag{14}$$

$$N_e = \frac{4N}{\mathrm{var}(k)+2} \tag{15}$$

$$\tag{16}$$

(c.f. **Wright1938-tv**).
Now, we look at this variance equation behaves with $t > 1$.

$$\mathrm{V}(x_1) = \mathbb{E}_1\left[(x_1 - x_0)^2\right] = \frac{x_0(1-x_0)}{2N}, \tag{17}$$

and

$$\mathrm{V}(x_2) = \mathbb{E}\left[(x_2 - x_0)^2\right] \tag{18}$$

$$= \mathbb{E}_1\left[\mathbb{E}_2\left[((x_2 - x_1) + (x_1 - x_0))^2 | x_1\right]\right] \tag{19}$$

$$= \mathbb{E}_1\left[\mathbb{E}_2\left[(\Delta x_2 + \Delta x_1)^2 | x_1\right]\right] \tag{20}$$

$$= \mathbb{E}_1\left[\mathbb{E}_2\left[{\Delta x_2}^2 | x_1\right]\right] + 2\mathbb{E}_1\left[\mathbb{E}_2[\Delta x_2 \Delta x_1 | x_1]\right] + \mathbb{E}_1\left[{\Delta x_1}^2\right] \tag{21}$$

$$= \mathbb{E}_1\left[\mathbb{E}_2\left[{\Delta x_2}^2 | x_1\right]\right] + \mathbb{E}_1\left[{\Delta x_1}^2\right] + \mathcal{C}_{1,2} \tag{22}$$

$$= \frac{p_0(1-p_0)}{2N}\left(1 - \frac{1}{2N}\right) + \frac{p_0(1-p_0)}{2N} + \mathcal{C}_{1,2} \tag{23}$$

Consequently, the variance of the neutral allele frequency changes each generation according to the probability of failing to coalesce each generation and the pairwise covariance terms $\mathcal{C}_{i,j}$ that build up due to associations with heritable fitness backgrounds. Note that the covariance term $\mathcal{C}_{1,2} = \mathbb{E}_1\left[\mathbb{E}_2[\Delta x_2 \Delta x_1 | x_1]\right] = 0$ under neutral evolution, and this simplifies to the well-known equation for variance in a Wright–Fisher population.

Now, we turn our attention to the case where there is heritable fitness variation, which leads to non-zero covariance terms $\mathcal{C}_{i,j}$. These terms emerge when $\mathrm{var}(k)$ has a heritable component of fitness that can be transmitted along with the neutral allele, thereby affecting the neutral allele's trajectory in later generations. If we partition the offspring number into heritable and non-heritable components, $k_i = 2f_i + \varepsilon_i$, then $\mathrm{var}(k_i) = 4V_h + V_n$. When there is heritable fitness variation across individuals, $V_h > 0$. Because the population size is assumed to be constant, the mean heritable fitness across individuals is constrained to be $\mathbb{E}_i(f_i) = 1$; this implies that the individual fitness values f_i are relative fitnesses as is standard in population genetics, in this case to the population

mean. Thus V_h is the squared coefficient of heritable fitness variation (this is often denoted as $C^2 = {}^{V_A}/_{\bar{w}^2}$, see **Crow1958-pc**; **Charlesworth1987-ab**; **Houle1992-ur**).

We can further decompose the K_1 term into $K_1 = S_1 + D_1$,

$$\begin{aligned}
\text{var}(\Delta x_1) &= \text{var}(S_1) + \text{var}(D_1) + \text{var}(H_1) && (24)\\
&\approx \frac{p_0(1-p_0)}{2N} V_h + \frac{p_0(1-p_0)}{2N}\frac{V_n}{4} + \frac{p_0(1-p_0)}{2N}\frac{1}{2}. && (25)\\
& && (26)
\end{aligned}$$

(c.f. **Santiago1995-hx** equation 11).

To see how the covariance terms accumulate, consider $\text{var}(x_3 - p_0)$. Note that $\mathbb{E}(S_t) = \mathbb{E}(D_t) = \mathbb{E}(H_3) = 0$, since which neutral allele we track is arbitrary, so by symmetry the expected change is zero. Then,

$$\begin{aligned}
\text{var}(x_3 - p_0) &= \mathbb{E}\left[(S_1 + D_1 + H_1 + S_2 + D_2 + H_2 + S_3 + D_3 + H_3)^2\right] && (27)\\
&= \mathbb{E}({S_1}^2) + \mathbb{E}({S_2}^2) + \mathbb{E}({S_3}^2) && (28)\\
&+ \mathbb{E}(S_1 S_2) + \mathbb{E}(S_1 S_3) + \mathbb{E}(S_2 S_3) && (29)\\
&+ \mathbb{E}({D_1}^2) + \mathbb{E}({H_1}^2) + \mathbb{E}({D_2}^2) + \mathbb{E}({H_2}^2) + \mathbb{E}({D_3}^2) + \mathbb{E}({H_3}^2) && (30)
\end{aligned}$$

The covariance terms $\mathbb{E}(S_i S_j)$ for $j > i$ represent the expected neutral allele frequency change from a neutral allele becoming associated with a fitness background in generation i and that fitness association persisting until generation j. However, the covariances terms $\mathbb{E}(S_j S_i)$ for $j > i$ are zero since associations in the future cannot affect the past (see p. 1041 of **Buffalo2019-qs**). Let us consider S_1 and S_2. We have,

$$\begin{aligned}
S_1 &= \frac{1}{2N}\sum_{i=1}^{N} x_{0,i}(f_{0,i} - 1) = \text{cov}(x_{0,i}, f_{0,i}) && (31)\\
S_2 &= \frac{1}{2N}\sum_{i=1}^{N} x_{1,i}(f_{1,i} - 1) = \text{cov}(x_{1,i}, f_{0,i}) && (32)
\end{aligned}$$

which are chance covariances (across the population, *not* evolutionary replicates) created by the random sorting of neutral alleles and fitness into individuals, *each generation*. These covariances are equivalent to the Robertson-Price equation (**Robertson1966-fs**; **Price1970-si**), which predict the change in neutral frequency due to heritable fitness (and likewise with the change due to non-heritable factors).

Considering the underlying haplotypes that lead to $x_{0,i}$ and $f_{0,i}$ can give us an equation for the dynamics of the $\mathbb{E}(S_i S_j)$ (for $j > i$) terms over evolutionary replicates. Let us partition the allele frequency and fitness by into the average of the paternal contributions. The neutral allele frequency per gamete is either 0 or 1, and we assume the fitnesses across gametes are additive. Then, as long as there is random mating, we can simplify the associations in a diploid by looking at a single gamete,

$$S_1 = \mathrm{cov}(x_{0,i}, f_{0,i}) = \mathrm{cov}\left(\frac{x'_{0,i} + x''_{0,i}}{2}, f'_{0,i} + f''_{0,i}\right) \tag{33}$$

$$= \frac{1}{2}\left(\mathrm{cov}(x'_{0,i}, f'_{0,i} + f''_{0,i}) + \mathrm{cov}(x''_{0,i}, f'_{0,i} + f''_{0,i})\right) \tag{34}$$

$$= \mathrm{cov}(x'_{0,i}, f'_{0,i} + f''_{0,i}) \tag{35}$$

$$= \mathrm{cov}(x'_{0,i}, f'_{0,i}) + \mathrm{cov}(x'_{0,i}, f''_{0,i}) \tag{36}$$

$$= S'_1 + S''_1 \tag{37}$$

which follows from symmetry, since it does not matter which gamete carrying the neutral allele we track (c.f. **Santiago1998-bs** p. 2107). The primes now indicate whether the selected locus is on the same gamete as the neutral allele (single prime), or the homologous gamete (double prime).

After meiosis, a fraction $1 - r$ of these associations between the tracked neutral neutral allele $x'_{0,1}$ and fitness background $f'_{0,i}$ persist, and a fraction r disassociate its currently coupled background and create an association with the homologous fitness background, $f''_{0,i}$. When linkage is tight, this latter term can be ignored as we will do here, though it does impact background levels of neutral diversity (**Santiago1995-hx**). Additionally, the fitness effects of the selected background may change in later generations, in complex ways (**Barton1986-yh**; **Turelli1990-kd**). **Santiago1995-hx** assume an equilibrium level of fitness variation V_h, with the fitness variance associated with a particular background decaying at a simple geometric decay at rate $1 - \kappa$ each generation, which is sufficient for background selection. An approximate rate of the decay can be worked out for the particular selective system. In general, this can be a much more complicated function; **Buffalo2019-qs** show that even fluctuating selection can be accommodated.

Now, the associations created in generation 1 and that persist in future generations as (ignoring S''_i terms) follow the pattern

$$S'_2 = S_1(1-\kappa)(1-r) \tag{38}$$

$$S'_3 = S_1(1-\kappa)^2(1-r)^2 \tag{39}$$

$$S'_4 = S_1(1-\kappa)^3(1-r)^3. \tag{40}$$

Then, the cumulative effect of the fitness associations created in generation 1 can be written as,

$$S_1 Q_t = S_1 + S'_2 + S'_3 + \ldots + S'_t \tag{41}$$

$$Q_t = 1 + \sum_{i=1}^{t} (1-r)^i (1-\kappa)^i \tag{42}$$

or for continuous t,

$$Q_t = 1 + \frac{(1-\kappa)(1-r)\left(1 - (1-\kappa)^t(1-r)^t\right)}{\kappa + (1-\kappa)r} \tag{43}$$

which as $t \to \infty$, converges to,

$$Q_\infty = \frac{1}{\kappa + r(1-\kappa)}. \tag{44}$$

This shows the long reach of heritable fitness factors in altering allele frequency through the generations. In terms like $\mathbb{E}(S_1 S_2)$, a fraction S_1' of the heritable allele frequency change S_2 were *previously* built associations between the neutral allele and a fitness background from S_1. As long as the direction of selection is the same, the expected change of this fraction of associations formed in the first generation to later generations 2 and 3 would be,

$$\mathbb{E}(S_1 S_2) = \mathbb{E}(S_1^2)(1-\kappa)(1-r) = \mathbb{E}(S_1^2)Q_2^2 \tag{45}$$

$$\mathbb{E}(S_1 S_3) = \mathbb{E}(S_1^2)(1-\kappa)^2(1-r)^2 = \mathbb{E}(S_1^2)Q_3^2. \tag{46}$$

With these covariance terms, we now have a full model for $\mathbf{var}(p_t - p_0)$. We will look at the case for $t = 2$. As shown in equation XXX, the second moment of every term is a function of the variance in neutral allele frequency in that generation, $x_t(1-x_t)$. This variance does not stay its initial value of $x_0(1-x_0)$; it decays each generation due to coalesce from the drift and selection processes modeled by this approach. The rate at which this decay happens is the effective population size each generation implied by this model. We can write,

$$\mathbf{var}(x_2 - p_0) = \mathbb{E}({S_1}^2) + \mathbb{E}(S_1 S_2) + \mathbb{E}({S_2}^2) \tag{47}$$

$$+ \mathbb{E}({D_1}^2) + \mathbb{E}({H_1}^2) + \mathbb{E}({D_2}^2) + \mathbb{E}({H_2}^2) \tag{48}$$

$$\frac{\mathbf{var}(x_2 - p_0)}{p_0(1-p_0)} = \frac{V_h}{2N_{e,1}} + \frac{V_h Q_2^2}{2N_{e,2}}\left(1 - \frac{1}{2N_{e,1}}\right) + \frac{V_h}{2N_{e,2}}\left(1 - \frac{1}{2N_{e,1}}\right) \tag{49}$$

$$+ \frac{V_n}{8N_{e,1}} + \frac{1}{4N_{e,1}} + \frac{V_n}{8N_{e,2}}\left(1 - \frac{1}{2N_{e,1}}\right) + \frac{1}{4N_{e,2}}\left(1 - \frac{1}{2N_{e,1}}\right). \tag{50}$$

In general, the evolution of the variance in allele frequency change in a system with heritable fitness due to a single locus r recombination fraction away from the neutral site is given by,

$$\frac{\mathbf{var}(x_t - p_0)}{p_0(1-p_0)} = \sum_{i=1}^{\infty}\left(\frac{V_h Q_i^2}{2N_{e,i}} + \frac{V_n}{8N_{e,i}} + \frac{1}{4N_{e,i}}\right)\prod_{t=1}^{i-1}\left(1 - \frac{1}{2N_{e,t}}\right). \tag{51}$$

$$\tag{52}$$

Since the probability of coalesce (or, equivalently, identity by descent) is proportional to the variance in allele frequency change (**Barton2000-zg**), this encodes the pairwise coalescent rate of the population through time under selection through the $N_{e,t}$ terms. Others have found that the genealogies under purifying selection can be characterized by such a time-dependent effective population size (**Nicolaisen2013-gv**). In the original derivation of **Santiago1995-hx**, they set each $N_{e,t}$ term to a single N and solve this summation in the limit to infinity giving,

$$\frac{\mathbf{var}(x_t - p_0)}{p_0(1-p_0)} = \frac{4V_h Q_\infty^2 + V_n + 2}{4(2N-1)}. \tag{53}$$

To relate this to an asymptotic effective population size, note that in an ideal Wright–Fisher population,

$$\mathrm{V}(x_t) = p_0(1-p_0)\left[1-\left(1-\frac{1}{2N_e}\right)^t\right] \tag{54}$$

such that if we look at the change in the variance in allele frequency change between adjacent generations t and $t-1$, we get an implied $N_{e,t}$

$$N_{e,t} = \frac{p_0(1-p_0) - \mathrm{var}(x_{t-1})}{2(\mathrm{var}(x_t) - \mathrm{var}(x_{t-1}))} \tag{55}$$

(cf. p. 1018, **Santiago1995-hx**).

Solving this for equation (53) gives the asymptotic effective population size,

$$N_e \approx \frac{4N}{4V_h Q_\infty^2 + V_n + 2} \tag{56}$$

(cf. equation 18, **Santiago1995-hx**). Levels of pairwise diversity, however, depend on the full distribution of pairwise coalesce times, or the $N_{e,t}$. While this would require solving the recursion defined by equations (51) and (55), **Santiago1998-bs** show that a reasonably good approximation follows from using a single effective population size for all timepoints, $N_{e,t} = N_e$, but use the time-indexed Q_t. This leads to,

$$N_{e,t} \approx \frac{4N}{4V_h Q_t^2 + V_n + 2} \tag{57}$$

(c.f. equation 15 of **Santiago1995-hx** and p. 1270 of **Santiago2016-mu**).

1.2 Effective Population Size Under Polygenic Selection

The results of the previous section assume only a single locus a recombination fraction r apart from the neutral locus is determining the coalesce rate. The model of **Santiago1998-bs** extends this to the case where fitness is polygenic. As they show in their original paper, this system can accommodate a variety of different equilibrium selection systems as long as expressions for V_h and $1-\kappa$ can be worked out. Other authors have extended similar models to more elaborate bbreeding structures (**Wray1990-zf**; **Woolliams1993-qo**).

First, we extend equation (57) to a polygenic system. Throughout, a multiplicative fitness model is assumed (i.e. independent effects, no epistasis), where the fitness of individual i is the product of their fitness contributions from each locus l, $w_{i,l}$, giving $w_i = \prod_{l=1}^n w_{i,l}$. Then, assuming mean fitness is one and independence between sites, the total fitness variation is multiplicative across the locus-level fitnesses,

$$V_h = \prod_{l=1}^{n}(1+V_{h,l}) - 1 \tag{58}$$

Similarly, the $Q_t^2 V_h$ term in the single-locus model is multiplicative under the polygenic model, as the change in site-specific fitness variances are assumed independent,

$$Q_t^2 V_h = \prod_{l=1}^{n} \left(1 + \frac{V_{h,l} Q_{t,l}^2}{2}\right) - 1 \tag{59}$$

where $Q_{t,l}$ is the cumulative impact of selection due to the selected site l (c.f. equation 4 **Santiago1998-bs**). The factor of $1/2$ is due to the fact that we are ignoring the chance that the fitness background on the homologous chromosome recombines onto the haplotype with the tracked neutral allele (i.e. due to the Q_t'' associations). In other words, only half the fitness variation in a diploid can stay associated with the neutral allele through the generations.

To simplify the derivation, assume $V_n = 2$ as under a Wright–Fisher model. Then, the effective population size is,

$$N_{e,t} \approx \frac{N}{V_h Q_t^2 + 1} \tag{60}$$

$$\approx \frac{N}{\prod_{l=1}^{n} \left(1 + \frac{V_{h,l} Q_{t,l}^2}{2}\right)} \tag{61}$$

$$\approx N \exp\left(-\sum_{l=1}^{n} \frac{V_{h,l} Q_{t,l}^2}{2}\right) \tag{62}$$

(c.f. **Santiago1998-bs** equation 4b). If we write each site l's contribution to the variance as a deviation from the average contribution, $V_{h,l}^2 = {}^1/_n V_h + \varepsilon_l$, then

$$N_{e,t} \approx N \exp\left(-\frac{1}{2} \sum_{l=1}^{n} \left(\frac{1}{n} V_h + \varepsilon_l\right) Q_{t,l}^2\right) \tag{63}$$

$$\approx N \exp\left(-\frac{V_h}{2n} \sum_{l=1}^{n} Q_{t,l}^2 + \sum_{l=1}^{n} \varepsilon_l Q_{t,l}^2\right) \tag{64}$$

Santiago and Caballero assume that over evolutionary replicates, as selected sites are randomly distributed, there is not systematic covariance between the fitness variation at a site and the decay of its association with the neutral site, e.g. $\mathbb{E}(\varepsilon_l Q_{t,l}^2) = 0$, and thus ignore the second sum term.

$$N_{e,t} \approx N \exp\left(-\frac{V_h}{2n} \sum_{l=1}^{n} Q_{t,l}^2\right) \tag{65}$$

$$\tag{66}$$

1.3 Continuous Approximation to the Segment Under Selection

Our model differs from that of **Santiago1998-bs** in that we do not integrate over the entire genome. In their 1998 model, Santiago and Caballero consider the impact of both sites under selection on

the same chromosome, as well as sites on independently assorting chromosomes. We only consider the contribution of a single segment under purifying selection, and the reduction from additional segments accumulate multiplicatively (similar to our implementation of the classic BGS model, XXX). Like their model, we imagine the reduction experienced by a focal neutral site in the middle of the segment. The implementation of our method differs slightly on though, see XXX.

Since $Q_{t,l}$ terms depend on l only through the recombination fraction between the neutral allele and the selected site, $r(l)$. Assuming a constant per-basepair recombination rate of r_{BP}, $r(l) = r_{\mathrm{BP}} l$ we can approximate the sum with an integral for the asymptotic Q_∞. The reduction experienced in the middle of a chromosome is twice the reduction experienced from each half,

$$N_{e,\infty} \approx N \exp\left(-\frac{V_h}{2L}\sum_{l=1}^{L} Q_{\infty,l}^2\right) \tag{67}$$

$$\approx N \exp\left(-\frac{V_h}{M}\int_0^{M/2} Q_\infty(r)^2 dr\right) \tag{68}$$

$$\tag{69}$$

where $M = n r_{\mathrm{BP}}$ is the total segment length in Morgans. Then, the total asymptotic effect of associations from all sites as Q_∞^2. Then,

$$Q_\infty^2 = \frac{2}{M}\int_0^{M/2} Q_\infty(r)^2 dr \tag{70}$$

$$= \frac{2}{M}\int_0^{M/2} \left(\frac{1}{1-(1-r)Z}\right)^2 dr \tag{71}$$

$$= \frac{2}{(1-Z)(2-(2-M)Z)} \tag{72}$$

Santiago and Caballero further approximate this; we can derive their approximation by setting $Z = 1-\kappa$ and doing a first-order Taylor series expansion around κ (since the loss in variance is presumed to be small). Then,

$$Q_\infty^2 = \frac{2}{M(1-Z)} + \frac{2(M-2)}{M^2} + \frac{2\left(M^2-4M+4\right)(1-Z)}{M^3} + \mathcal{O}(\kappa^2) \tag{73}$$

Santiago and Caballero's approximation only keeps this first term, e.g. their $Q_\infty^2 \approx 2/(M(1-Z))$, which is only accurate in the domain $M > 0.2$.

$$N_{e,\infty}^{(SC98)} \approx N \exp\left(-\frac{V_h}{(1-Z)M}\right) \tag{74}$$

(c.f. **Santiago1998-bs** equation 8). For small M, the small κ approximation strongly deviates from the full equation; thus, throughout, we use the unapproximated version,

$$N_{e,\infty} \approx N \exp\left(-\frac{V_h}{(1-Z)(2-(2-M)Z)}\right). \tag{75}$$

For the reasons described earlier, a single asymptotic $N_{e,\infty}$ cannot fully characterize pairwise coalescent rates in all cases, and thus neutral diversity levels depend on the full $N_{e,t}$, for which we integrate equation (43) to get,

$$N_{e,t} \approx N \exp\left(-\frac{V_h}{M}\int_0^{M/2}\left(1+\frac{(1-k)(1-r)\left(1-(1-k)^t(1-r)^t\right)}{k+(1-k)r}\right)dr\right). \tag{76}$$

1.4 The Variance Dynamics

The heritable variance in fitness V_h associated with the tracked neutral allele does not remain constant through the generations, but decays due to selection and drift, and is maintained by the input of new mutations. For an arbitrary polygenic selective system, the dynamics are extraordinarily complicated due to the complexity of multilocus selection, and the dynamics of the trait response to selection depend on the variance, which in turn depends on higher moments of the fitness distribution. We discuss the simplification of Santiago and Caballero's model here; hereafter, we assume that $V_h = V_A$, the genetic variance in additive fitness put on a relative scale such that $\mathbb{E}_i(f_i) = 1$.

Each generation, the genetic variance changes due to changes in the allele frequencies at selected sites and linkage disequilibria between these sites. These are typically expressed as recursions, but certain models permit these changes to be expressed as a proportional reduction (e.g. truncation selection and stabilizing selection models; **Keightley1988-eq** p. 36, and **Walsh2018-bt** p. 557). Santiago and Caballero's model follows this approach, assuming the variance decays due to selection and drift at a constant rate $Z := 1-\kappa$ through the generations, such that the variance in the next generation excluding mutation is $V_A' = (1-\kappa)V_A$ and $\Delta V_A = -\kappa V_A$.

With mutation, $\Delta V_A = V_m - \kappa V_A$. The model of Santiago and Caballero assumes that under the long-run equilibrium, $\Delta V_A = 0$, such that $\kappa V_A = V_m$, and thus the rate of loss is $\kappa = {}^{V_m}/_{V_A}$. Then the variance for the dynamics simplify to,

$$V_A' = \left(1-\frac{V_m}{V_A}\right)V_A \tag{77}$$

$$Z := \frac{V_A'}{V_A} = 1-\kappa = 1-\frac{V_m}{V_A} \tag{78}$$

(c.f. **Santiago1998-bs** equation 9).

1.5 Effective Population Size Under Strong Background Selection

Under strong background selection, multiple segregating deleterious sites contribute to fitness variation. BGS models assume a multiplicative fitness model, such that the fitness of an individual is

$$w = (1-s)^n \tag{79}$$

where n is the number of deleterious mutations this individual carries.

Classic background selection theory assumes an infinite population size, such that mutation frequencies are at their mutation-selection equilibrium and the expected number of mutations across

individuals is $\mathbb{E}(n) = \mu/s L$ where L is the total number of basepairs that can be mutated (the mutational target size). Under a multiplicative fitness function in an infinite population, selection cannot generate linkage disequilibria (**Turelli1990-kd**). Thus the classic strong background selection model assumes additive genetic fitness is equal to the additive genic fitness (which excludes the contribution of covariance between selected sites),

$$V_{\text{BGS}} := \text{V}(w) = 2s^2 \sum_{l=1}^{L} p_l(1-p_l) \tag{80}$$

$$= 2Ls^2 \frac{\mu}{s}\left(1 - \frac{\mu}{s}\right) = Us + \mathcal{O}(\mu^2) \approx Us. \tag{81}$$

since under mutation-selection balance $p_l = \mu/s$ and U is defined $U := 2L\mu$, the deleterious mutation rate per diploid genome, per generation. They then set $V_A = V_{\text{BGS}}$, and note that $V_m \approx Us^2$, since mutations have frequency of $1/2N$ and each increase the variance by a factor of s^2. Note that because the classic BGS model assumes deterministic selection dynamics (i.e. an infinite population), the equilibrium Z given by equation (77) excludes the reduction due to drift. The reduction factor under the strong BGS model is,

$$Z_{\text{BGS}} := 1 - \frac{Vm}{V_{\text{BGS}}} = 1 - \frac{Us^2}{Us} = 1 - s \tag{82}$$

(c.f. **Santiago1998-bs** equation 10). An alternate derivation can be found by working out the deterministic reduction in variance from the single-locus dynamics.

Now, we can use these values for Z_{BGS}, V_m, and V in equation (75). Through this alternate quantitative genetics derivation, we arrive at the same equation as classic BGS theory,

$$N_{e,\infty}^{\text{BGS}} = N \exp\left(-\frac{U}{2s + M - Ms}\right). \tag{83}$$

This matches equation 8 of **Hudson1995-xc** except for the Ms term and equation 10 of **Nordborg1996-nq** except for the $2s$ term.

1.6 Effective Population Size Under Weak and Strong BGS

With the possibility of weak selection, deleterious alleles can drift up in frequency and fix. Consequently, the fitness distribution is no longer stationary, as the set of haplotypes without any deleterious mutations can be lost due to drift. This changes the shape of the fitness distribution, and thus the variance parameter that our model relies upon (**Gessler1995-hz**; **OFallon2010-my**; **Good2013-lp**; **Haigh1978-gt**; **Higgs1995-xc**). So far, finding an explicit equation for the variance has been difficult due to the "moment-closure problem": the variance depends on higher moments such as the skew, which in turn depends on higher moments. While the weak selection regime is composed of an interaction of mutation, selection, drift, and linkage processes and the dynamics that differ qualitatively based on parameter combinations. In their approximation, **Santiago2016-mu** consider the modified additive genic variance due to the fixation of weakly deleterious mutations. This fixation process is related to the Muller's Ratchet, the rate R of fixation

of deleterious mutations per generation per region, since each fixation directly reduces the additive genic variance of the fitness distribution. However, in general, the rate of this ratchet is unknown since it depends on the fitness variance (and thus higher moments of the fitness distribution).

In their 2016 paper, Santiago and Caballero derive a general equation for the fitness variance that accounts for the reduction due to the fixation of weakly deleterious alleles (i.e. the rate of Muller's ratchet). There derivation for the variance uses a heuristic approach based on Fisher's Fundamental Theorem of Natural Selection (equation 2, **Santiago2016-mu**, equation 15A **Garcia-Dorado2007-jj**). A more formal derivation follows from **Higgs1995-xc** (though they assume a haploid asexual system, the end result is the same up to a factor of two). In their work, they derive equations for the change per generation for the moments and cross-moments of the fitness distribution under multiplicative selection in a haploid model. They find, the haploid rate of the ratchet is $r = u - sv_n$ (equation 13.3, **Higgs1995-xc**), where we use lowercase to distinguish between haploid and diploid models. Here, v_n is the variance in the *number* of deleterious mutations; as long as selection is not too strong, the haploid additive variance is $v_A \approx s^2 v_n$. Substituting this for v_n and rearranging, the haploid fitness variance is related to the $v_A = us - rs$, which is identical to Santiago and Caballero's (**Santiago2016-mu**) equation 2, noting that the waiting time between fixations (i.e. ratchet clicks) is $T = 1/r$. Intuitively, the fitness variance Us is reduced by the rate at which deleterious mutations fix and remove fitness variation from the population.

Now, note that under a diploid model, the fitness variance is twice that as the haploid model, so $V_A = 2v_A$. Then, the fitness variance is,

$$V_A = Us - 2Rs. \tag{84}$$

where $U = 2u$ and $R = r$ is the diploid rate of the ratchet, which depends on higher moments of the fitness distribution, but only differs from r through a rescaling of the population size. Note that if selected sites are sufficiently deleterious that they cannot fix (or the population is infinite), $R = 0$ and the additive fitness variance is identical to that under strong background selection, $V_A = Us = V_{\text{BGS}}$ (equation (80)). Then, the fixation probability of a new deleterious mutation with additive effects ($1 - s$ in the heterozygote, $1 - 2s$ in the homozygote) in a diploid population is,

$$p_{\text{fix}} = \frac{1 - \exp(2N_e s/N)}{1 - \exp(4N_e s)} \approx \frac{2N_e s}{N\left(\exp(4N_e s) - 1\right)} \tag{85}$$

(c.f. **Durrett2008-ql**, equation 7.21; **Kimura1957-rk**, equation 5.6). Then, Santiago and Caballero argue that the rate of the ratchet is the inverse of the averaging waiting time until a fixation, $T \approx 1/(UNp_{\text{fix}})$ (note that since U is the *diploid* mutation rate, the total population mutation rate is UN per generation), so

$$T \approx \frac{\exp(4N_e s) - 1}{2UsN_e} \tag{86}$$

Since the ratchet rate depends on N_e, which in turn depends on the rate of the ratchet, we write the ratchet, or substitution, rate per basepair per generation as,

$$R(N_e) := \frac{2UsN_e}{\exp(4N_e s) - 1}. \tag{87}$$

Under weak and strong background selection the variance input of mutations V_m is the same, and using the V_A derived above, the decay in variance is

$$Z_{\text{WS}} = 1 - \frac{V_m}{V_A} = 1 - \frac{Us^2}{(U - R(N_e))s} = 1 - \frac{Us}{U - R(N_e)}. \tag{88}$$

At equilibrium, the N_e that determines fixation probability is the asymptotic $N_{e,\infty}$. Consequently, Santiago and Caballero argue that the asymptotic reduction in effective population size uses the Q_∞^2 and is given by the following system of non-linear equations,

$$R(N_{e,\infty}) = \frac{4UsN_{e,\infty}}{\exp(4N_{e,\infty}s) - 1} \tag{89}$$

$$N_{e,\infty} \approx N \exp\left(-\frac{V_A}{(1-Z)(2-(2-M)Z)}\right) \tag{90}$$

The solution to these equations is the asymptotic equilibrium $\widetilde{N}_e$ and equilibrium ratchet rate $\widetilde{R}$. This sets the equilibrium variance, $\widetilde{V} = Us - 2\widetilde{R}s$, or $\widetilde{V} = (U - 2\widetilde{R})s$. Note that because $p_{\text{fix}} \leq 1/2N$ for deleterious to neutral variation, $2R \leq U$, or the diploid mutation rate must be greater than twice the per-basepair per-generation substitution rate for equilibrium variance $\widetilde{V} > 0$ for $s > 0$.

1.7 Steady-state Equilibrium Genic Variance

We also find that Equation (**??**) (and Supplementary Materials Equation 84 above) can also be found through Kimura's diffusion models with a flux of mutations into infinite, discrete sites (**Kimura1969-jw**). This model integrates the Kolmogorov backwards equation over time and forward variable to find the expected value of a functional of the initial frequency. Using the functional $I(p) = s^2 p(1-p)$ for genic variance given mutations start at frequency p, one finds that the steady-state genic variance for $p = 1/2N_e$ is

$$\begin{aligned} V_a^{SS} &= 2UNs\left(\frac{1}{2N_e} - p_F(N_e, s)\right) \\ &= \left(\frac{UN}{N_e} - 2UNp_F(N_e, s))\right)s \\ &\approx (U - 2R)\, s \end{aligned} \tag{91}$$

which is approximately Equation (**??**).

1.8 Heterozygosity and B values Under Weak and Strong Selection

While the selection dynamics are set by the asymptotic Q_∞^2 that determines the equilibrium $\widetilde{N}_e$ and $\widetilde{R}$, heterozygosity is determined by the sequence $\widetilde{N}_{e,t}$, where each term is given by (76). We experimented with calculating this full sum in our methods, but it seemed to make little difference in our application. Additionally, it was too costly to calculate computationally for genome-wide calculations of the B' map.

1.9 Numeric methods for calculating B

Depending on the number of features, this can be incredibly memory intensive.

1.10 Optimization

Optimization in non-linear regression is notoriously difficult both in general (Bates) and in the specific problem of estimating the effects of genome-wide selection (**Murphy2022-sj**). We experimented with a variety of global and local optimization approaches and parameterizations, but found that

2 Human Genomic Data

2.1

In order to try to ensure accurate estimation of pairwise diversity, which is a ratio estimator that is sensitivity to its denominator, we use the complete per-basepair genotype calls (gVCF) produced by Illumina's DRAGEN pipeline (**Illumina˙Inc2020-dk**). The original samples were from 178 Yoruban individuals sequenced to 30x by the New York Genome Center (**Byrska-Bishop2022-tn**). This allows for filtering to be applied to the entire genome at once, rather than just variants, so the denominator does not need to be estimated separately.

The full list of samples is available in TSV format in the GitHub repository (`data/h1kg/yri_samples.tsv`).

2.2 Filtering gVCFs

gVCFs were filtered using a custom Python tool, `gvcf2counts.py` (in `tools/gvcf2counts.py`), which reads the gVCFs, filters them according to the criteria below, and outputs a Numpy `.npy` file of reference and alternative allele counts for each chromosome (hereafter, "allele counts matrices").

Genotypes are included in the allele count if and only if:

1. The variant call is set to `PASS` in the VCF.
2. The `QUAL > 50`.
3. The `GQ > 30` (or `RGQ > 30` for invariant sites).

Because the data underlying the counts files are per-basepair resolution gVCFs, each chromosome's allele counts matrix is of size $l \times 2$, where l is the total chromosome length. Basepairs that fail these filtering requirements lead to a row of zero counts, e.g. no observed reference and

alternative allele counts, and thus do not effect the data that goes into the binomial likelihood or π estimates used in figures.

2.3 Site-based Filtering of Counts

The allele counts matrices include many basepairs that may have allele counts that pass the genotype call filters, but are still need to be filtered out because the region of the genome may produce unreliable estimates. The following filters are applied based on masking regions:

1. **Non-accessible regions**: masks out centromeres (`acen` entries in `cytoBand.txt`), with 5Mbp padding on either side. The file of passing masks is `data/annotation/no_centro.bed`.

2. **Reference masking**: soft and hard-masked regions in the human GRCh38 reference genome are also masked. Soft-masked regions were determined by Ensembl (XXX), which uses Repeat Masker (XXX) and Dust (XXX).

3. **"Putatively" neutral regions**:

   Additionally, for fitting our likelihood and estimating observed pairwise diversity, we only consider . This masks out phastCons regions (from `phastConsElements100way.bed.gz`) and Ensembl gene regions (from annotation file `Homo_sapiens.GRCh38.107.chr.gff3.gz`). While introns are possibly under some weak selection, they collectively make up nearly 40% of the human genome and are included so genome-wide diversity can be estimated more precisely (possibly at the expense of some bias).

These files are all produced by the Snakemake file `data/annotation/Snakefile`. Note that the levels of accessible putatively neutral bases

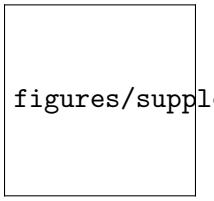
figures/supplementary/chr1_diversity_filtering.pdf

Figure 1: Estimates of chromosome 1 YRI diversity in non-overlapping megabase windows, under different filtering criteria. The filtering criteria are, (1) "neutral, accessible" which includes only putatively neutral sites, and ignores regions masked as inaccessible, (2) "accessible" which is only ignores sites masked as inaccessible, and (3) "no filtering" which uses all available data. Note that filtering changes the absolute level of diversity, but has more minor effects on regional patterns of diversity at the megabase scale.

2.4 Data Summary Matrices

Our underlying data for all likelihood and pairwise diversity estimates is the allele count matrix $\mathbf{C}$ with dimensions $L \times 2$, where L is the chromosome length. This is transformed to a pairwise summary matrix with identical dimensions, $\mathbf{Y}$. The first column of $\mathbf{Y}$ is the number of pairwise comparisons between chromosomes that are identical, and the second column is the number that

chrom	accessible	neutral	both
chr1	42.4	64.4	22.9
chr2	46.8	58.7	23.7
chr3	44.1	62.9	24.8
chr4	44.0	59.3	21.5
chr5	44.1	58.0	21.4
chr6	45.0	58.0	22.1
chr7	44.4	65.8	26.4
chr8	43.8	61.2	23.1
chr9	39.2	64.5	20.3
chr10	44.6	62.8	25.3
chr11	42.7	63.5	23.4
chr12	41.7	62.7	22.7
chr13	39.7	62.0	18.3
chr14	37.6	65.4	19.0
chr15	37.0	69.4	20.3
chr16	39.4	63.8	20.9
chr17	40.4	64.0	22.3
chr18	40.9	59.8	19.8
chr19	30.9	69.8	18.0
chr20	36.9	61.6	19.8
chr21	31.6	68.6	18.2
chr22	26.9	75.0	16.4

are different. Both of these columns are combinatoric summaries of the raw allele counts matrix needed for the binomial likelihood and pairwise diversity estimates. Let $[c_1, c_2]$ be a row of $\mathbf{C}$ for basepair l (the l index is omitted for clarity), and $n = c_1 + c_2$. Then, the (1) total number of pairwise combinations of chromosomes n_T, (2) the number of pairwise with identical alleles n_S, and (3) the number of pairwise combinations with differing alleles n_D are respectively,

$$\begin{aligned} n_T &= \frac{n(n-1)}{2} \\ n_S &= \binom{c_1}{2} + \binom{c_2}{2} \\ n_D &= n_T - n_S \end{aligned}$$

which would be stored in row $\mathbf{Y}_l = [n_S, n_D]$. Note that the per-site $\mathbf{Y}$ handles non-polymorphic sites and missing data. Non-polymorphic sites have $n_S = \binom{n}{2}$ and $n_D = 0$, and missing data has $n_S = n_D = 0$.

2.5 Pairwise Diversity Estimates

The pairwise diversity at site l across the n sampled chromosomes can be calculated from row l of the $\mathbf{Y}$ matrix as follows,

$$\pi_l = \frac{n_D}{n_T} \tag{92}$$

which is identical to the more common expression of this estimator,

$$\pi_l = \frac{2}{n(n-1)} \sum_{i<j}^{n} k_{i,j} \tag{93}$$

where $k_{i,j}$ is 1 if the alleles at this site differ at site l, and 0 otherwise.
There are three ways to aggregate π_l across all sites. The first is,

$$\pi^{(1)} = \frac{1}{L} \sum_{i=1}^{L} \frac{n_{D,i}}{n_{T,i}} \tag{94}$$

which if the number of samples across loci is constant, simplifies to an unweighted average across sites. Second, one can take a weighted average, with weights determined by the total number of samples present at a site,

$$\pi^{(2)} = \frac{1}{\sum_{i=1}^{L} n_i} \sum_{i=1}^{L} n_i \frac{n_{D,i}}{n_{T,i}} \tag{95}$$

Third, one can weight by the number of pairwise comparisons at a site, $n_{T,i}$, rather than total number of samples, n_i, which leads to a ratio of sums,

$$\pi^{(3)} = \frac{\sum_{i=1}^{L} n_{D,i}}{\sum_{i=1}^{L} n_{T,i}}. \tag{96}$$

We predominantly use the estimator $\pi^{(3)}$, as it corresponds to how we summarize the matrix $\mathbf{Y}$ across windows for our likelihood. All methods have mean squared errors and biases very close to one another (TODO).

Note, however, that estimates of pairwise diversity often condition on the accessible bases, and thus treat this as fixed. However, the number of accessible bases varies across the chromosome; this can lead to a source of apparent bias during block-bootstrap estimates of uncertainty. In this case, pairwise diversity is a ratio estimator, and is thus biased, since by Jensen's inequality $\mathbb{E}(y/x) \geq \mathbb{E}(y)/\mathbb{E}(x)$ for random variables x and y.

2.6 Window-based Summaries and filtering

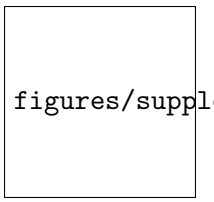

Figure 2: The distribution of diversity across the genome at the megabase scale, with outliers trimmed. The blue line is the normal CDF, fit with MLE parameters XXX. The inset figure is the untrimmed data, with trimmed points shown in red.

The likelihood method is fit to non-overlapping binned summaries of the allele counts matrix. Based on exploratory data analyses, some bins were outliers and excluded (XXX).

1. **Fraction of inaccessible sites per window**. `mask_inaccessible_bins_frac`
2. **Outliers**: Based on exploratory analysis, there were some regions with very high diversity. The 0.05% right tail was excluded.

3 Additional Results

3.1 Model Comparisons

4 Table of R^2 Values

5 Derivation of Approximate Drift R^2

Our approach models the observed megabase-scale diversity y_i in window i as the predicted level under our negative selection model, $\pi_0 B_i$, plus some random residual error ε_i,

$$y_i = \pi_0 B_i + \varepsilon_i \tag{97}$$

Table 1: R^2 and mutation rate estimates for all models. XXX note about repeats

Model	**Track type**	**Pop.**	**B'** $R^2_{\mathbf{LOO}}$	**B'** $R^2_{\mathbf{IS}}$	**B** $R^2_{\mathbf{IS}}$	**B'** $\hat{\mu} \times 10^{-8}$	**B** $\hat{\mu} \times 10^{-8}$
phastcons>CDS>genes	sparse	YRI	**68.45**	68.16	65.23	1.8636	0.2975
phastcons>CDS>genes	full	YRI	68.19	68.12	63.67	1.7082	0.1666
CADD 6%	full	YRI	68.08	68.08	62.96	1.4612	0.1721
CADD 6%	sparse	YRI	68.04	68.03	68.01	2.0912	2.0716
CADD 8%	full	YRI	66.98	67.11	63.61	1.0126	0.1713
CADD 8%	sparse	YRI	66.67	67.01	66.98	1.5688	1.5473
CDS>genes>phastcons	full	YRI	64.90	65.51	61.23	4.1951	0.1620
CDS>genes>phastcons	sparse	YRI	64.90	65.51	63.16	4.1795	0.3070
phastcons>CDS>genes	sparse	CEU	**61.98**	63.02	60.02	2.0194	0.3036
phastcons>CDS>genes	full	CEU	61.78	63.01	58.57	2.3499	0.1679
CADD 6%	full	CEU	61.74	62.94	57.42	1.5365	0.1739
CADD 6%	sparse	CEU	61.66	62.86	62.84	2.1215	2.1084
CADD 8%	full	CEU	60.80	62.12	57.95	1.1572	0.1742
CADD 8%	sparse	CEU	60.59	62.06	62.04	1.6030	1.5876
CADD 6%	full	CHB	59.62	61.31	55.31	1.5083	0.1727
CADD 6%	sparse	CHB	59.54	61.21	61.20	2.1242	2.1170
phastcons>CDS>genes	sparse	CHB	**59.44**	61.08	57.88	2.1982	0.3060
phastcons>CDS>genes	full	CHB	59.29	61.08	56.71	2.5573	0.1694
CADD 8%	full	CHB	58.88	60.63	55.86	1.1128	0.1733
CADD 8%	sparse	CHB	58.65	60.57	60.55	1.6098	1.6001
CDS>genes>phastcons	sparse	CEU	58.58	60.30	58.41	4.1041	0.3110
CDS>genes>phastcons	full	CEU	58.58	60.30	56.75	4.1001	0.1648
CDS>genes>phastcons	sparse	CHB	56.48	58.73	56.60	3.7195	0.3136
CDS>genes>phastcons	full	CHB	56.48	58.73	54.95	3.7552	0.1654

where each $B_i = f(X_i|\Psi)$ for some data X_i (i.e. the recombinaton map and annotation model) and parameters Ψ. To compare our models and assess model goodness-of-fit, we use the out-sample R^2_{LOO} which is calculated by fitting the genome-wide model leaving one chromosome out, and calculating the residual variance between predictions and observed values for this out-sample chromosome.

In general, R^2 is calculated for a model fit across windows as,

$$R^2 = 1 - \frac{V_{\text{res}}}{V_{\text{tot}}} \tag{98}$$

where V_{res} is the mean squared error across windows and V_{tot} is the total variance in the predicator. Suppose we have estimates $\widehat{\pi}_0\widehat{B}_i$ for $\pi_0 B_i$ from our model. Then,

$$V_{\text{res}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\pi}_0\widehat{B}_i)^2 \tag{99}$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2}_{\text{irreducible error}} + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}(\pi_0 B_i - \widehat{\pi}_0\widehat{B}_i)\right)^2}_{\text{bias squared}}. \tag{100}$$

The total variance in the predictor is,

$$V_{\text{tot}} = V_{\text{res}} + V_{\text{model}} + 2\widehat{\pi}_0 \operatorname{cov}_i(\varepsilon_i, \widehat{B}_i) \tag{101}$$

where

$$V_{\text{model}} = \frac{\widehat{\pi}_0^2}{n}\sum_{i=1}^{n}\left(\widehat{B}_i - \frac{1}{n}\sum_{i=1}^{n}\widehat{B}_i\right)^2. \tag{102}$$

Generally, fitting procedures act to minimize the squared deviation between predictors and observed values, which drives $\operatorname{cov}_i(\varepsilon_i, \widehat{B}_i)$ to zero XXX. **Murphy2022-sj** suggest XXX

Thus, our R^2 depends on (1) the bias of our predictors, (2) the covariance between residuals and predictors, and (3) the irreducible error due to variance in diversity within windows. However, it is of interest to approximate the irreducible error under theoretic models that assume all irreducible error is due entirely to the variance of coalescence times in a window, assuming that all of the effects of negative selection on the variance can be thought of as a local reduction in the effective population size to $B_i N_e$ and demographic effects amount to a simple rescaling of N_e. This also assumes mutation rates are constant across the genome, and do not contribute to the irreducible variance. We call this theoretic goodness-of-fit under drift R^2_{drift}, and it provides a rough approximation of the irreducible error in our model due to "coalescence noise" around the modeled reductions due to negative selection.

Then,

$$\mathbb{E}(\varepsilon_i^2) = \operatorname{var}(\pi_i) \tag{103}$$

since the mean residual is zero. The variance in pairwise diversity $\text{var}(\pi_i)$ is equivalent to the variance in the number of segregating sites for a sample of two, $\text{var}(\pi) = \text{var}(S_2)$. The variance in coalescence times in a window of width w basepairs with population-scaled recombination rate $\rho = 4B_i N_e r w$ is,

$$\text{var}(S_2) = \theta + \theta^2 \frac{2}{\rho^2} \int_0^\rho (\rho - x) f_2(x) dx \tag{104}$$

(**Wakeley2009-ua**) where,

$$f_2(\rho) = \frac{\rho + 18}{\rho^2 + 13\rho + 18}. \tag{105}$$

Since this is a rough approximation, we set $\theta = w\pi_0 B_i$ and fix $\rho = w\gamma\pi_0 B_i$ where γ is the ratio of recombination to mutation rates, r/μ. In humans, $\gamma \approx 1$, but we explore different ratios to assess sensitivity of this calculation (Supplementary Figure **??**).

This back-of-the-envelope calculation finds that the upper bound of R^2 would be around 67% for Yoruba, and 64% for European and Han Chinese models (Figure 3 This rough approximation assumes constant demography, which for bottlenecked out-of-Africa populations assumes that the variance in coalescence times is determined entirely by reducing the effective population size. However, the variance in coalescence times is *higher* in bottlenecked populations (CEU and CHB) compared to Yoruba, which we confirm with simple simulations using the `OutOfAfrica_3G09` model from `stdpopsim` (**Gutenkunst2009-pg**; **Adrion2020-cf**). This higher than expected variance would inflate SS_{res}, which would decrease R^2_{drift}, bringing it closer to the observed R^2_{LOO}. A more accurate approximation of the R^2_{drift} could be found with simulations with more realistic demography and varying coalescence rates along the genome.

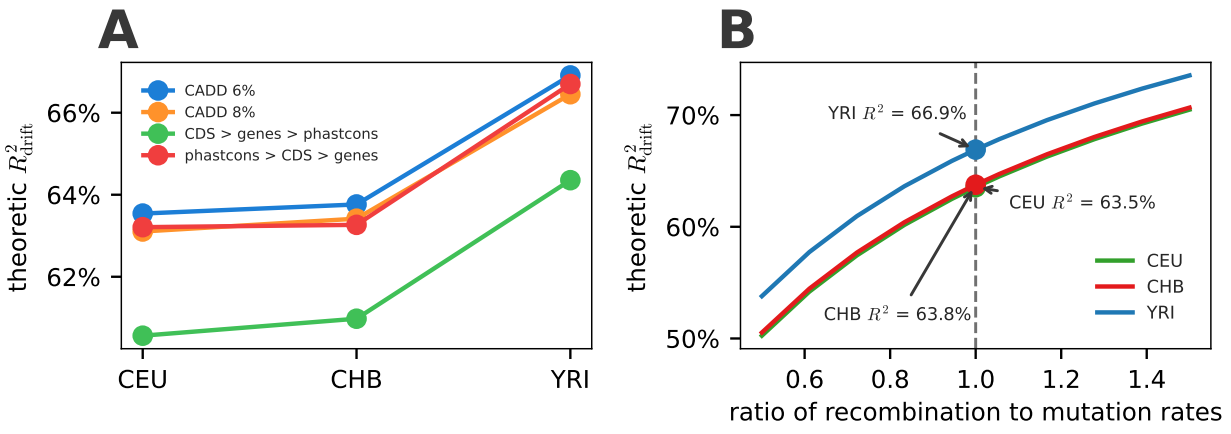

Figure 3: (A) the theoretic R^2_{drift} using different plugin estimators for B_i under different models. This shows that among the best-fitting models, the predicted R^2_{drift} varies little; most variation is among populations due to differing π_0. (B) The R^2_{drift} across different populations, for varying ratio of recombination to mutation rates (x-axis).

6 Predicted Rate of Fitness Change

Each fixation of a mutant with fitness effect $-s$ reduces the population fitness to $1-2s$ since all individuals are homozygotes for the deleterious mutation and we assume additive effects at a site. For each segment l, the maximum likelihood parameter estimates of our model imply a prediction of the deleterious substitution rate per segment $\widehat{R}_{l,i}$ for each selection coefficient s_i in the m_s-element DFE grid (according to which feature class it belongs to).

Since we assume multiplicative fitness effects, the log population fitness (considering only fixed sites) for segment l in the next generation is,

$$\log(L_l) = \sum_{i=1}^{m_s} \log(1-2s_i)\widehat{R}_{l,i}. \tag{106}$$

Since we assume multiplicative fitness, the total genome-wide log population fitness in the next generation is

$$\log(L) = \sum_{l} \log L_l. \tag{107}$$

This amounts to a predicted population fitness loss per generation of

$$\begin{aligned} \widehat{\Delta}_W &= 1 - L \\ &= 1 - \exp\left(\sum_{l}\sum_{i=1}^{m_s} \log(1-2s_i)\widehat{R}_{l,i}\right). \end{aligned} \tag{108}$$

We calculate this value for all segments within a feature class, and then sum over all features to arrive at our genome-wide estimate. We use a parametric bootstrap to carry forward the estimation uncertainty to the predicted fitness loss per generation. Briefly, we do this by sampling new estimates of the DFE per selection coefficient and feature from a normal distribution centered on the ML point estimate and with the jackknife standard deviation.

7 Likelihood

Our model is essentially a generalized nonlinear model with a binomial link function. This is the form used by **Elyashiv2016-vt** and **Murphy2022-sj**. These models fit the observed number of pairwise nucleotide differences in genomic windows to the expected pairwise diversity under some evolutionary model. We only consider BGS models here, so the mean function for position x is the product of the proportion by which BGS reduces diversity at position x, $B(x)$, and genome-wide neutral diversity π_0,

$$\pi(x) = B(x, \Phi)\pi_0 \tag{109}$$

Φ is the set of BGS parameters (i.e. the DFE for each feature type and mutation rate), and $\pi_0 = 4N_e\mu$ is determined by the genome-wide drift-effective population size N_e, set by only

reproductive and demographic processes. Regional mutation rate heterogeneity can be accounted for with a regional mutation rate scaling function $m(x)$, $\pi(x) = B(x, \Phi)m(x)\pi_0$, but we do not explore mutation-rate heterogeneity, as it is unclear how to differentiate variance in mutation rate from the substitution rate heterogeneity we find across the genome.

The likelihood of the set of parameters $\Psi = \{\pi_0, \Phi\}$ can be written in the form of,

$$\log \mathcal{L}(\Psi) = \sum_{v \in \mathcal{V}} \sum_{i \neq j \in \mathcal{S}} \log(P(O_{i,j}(v)|\Psi)) \tag{110}$$

(c.f. **McVicker2009-ax**; **Elyashiv2016-vt**; **Murphy2022-sj**) where $\mathcal{V}$ is the set of putatively neutral sites, $\mathcal{S}$ is the set of samples, and Ψ are the BGS parameters. The indicator variable $O_{i,j}(v)$ is 1 if samples i and j are different at putatively neutral site v, and zero otherwise. While the theoretic $\pi(v)$ gives the average number of pairwise differences, for small values, this is approximately the heterozygosity probability, so we can write

$$P(O_{i,j}(v)|\Psi) = \begin{cases} \pi(v), & O_{i,j}(v, \Psi) = 1 \\ 1 - \pi(v), & O_{i,j}(v, \Psi) = 0 \end{cases}. \tag{111}$$

(c.f. **Elyashiv2016-vt**).

As in Section 2.5, the number of pairwise differences and the total number of pairwise comparison at a site are sufficient statistics for the likelihood. Then, the binomial log-likelihood for the data at site v is,

$$\ell_v(\Psi) = \log(\pi(v, \Psi))n_{\mathrm{D},v} + \log(1 - \pi(v, \Psi))n_{\mathrm{S},v} \tag{112}$$

7.1 The scale of processes

We can observe $\widehat{\pi}(x)$ at a per-basepair resolution. However, for a variety of reasons, we do not want to fit the composite likelihood model to the per-basepair scale of data. First, this would be computationally infeasible. Second, the mean function $\pi(x)$ varies on a natural scale that is itself a free parameter of the model. Our model can be written as,

$$\ell(\Psi, h) = \sum_b \sum_{v \in \mathcal{V}_b} \ell_v(\Psi) \tag{113}$$

$$= \sum_b \left[\log(\bar{\pi}(b, \Psi)) \sum_{v \in \mathcal{V}_b} n_{\mathrm{D},v} + \log(1 - \bar{\pi}(b, \Psi)) \sum_{v \in \mathcal{V}_b} n_{\mathrm{S},v} \right] \tag{114}$$

$$= \sum_b [\log(\bar{\pi}(b, \Psi))Y_{\mathrm{D},b} + \log(1 - \bar{\pi}(b, \Psi))Y_{\mathrm{S},b}] \tag{115}$$

$$\tag{116}$$

where h is the bandwidth or window size, b the bin index for windows of width h, $\mathcal{V}_b$ is the set of putatively neutral sites in bin b, $\bar{\pi}(b|\Psi)$ are the average diversity in bin b, and $Y_{\mathrm{S},b}$ and $Y_{\mathrm{S},b}$ are the sums across putatively neutral sites within a bin. Note that by binning, we sum the pairwise

summaries of the data $\mathbf{Y}$ across sites, so the likelihood across bins is naturally weighted by the quantity of observed data.

This model corresponds to a binomial likelihood for the observed data summarized at genomic scale h. Thus an alternative way to express this model is as,

$$Y_{\mathrm{D},b} \sim \mathrm{Binom}(\bar{\pi}(b,\Psi), Y_{\mathrm{D},b} + Y_{\mathrm{S},b}). \tag{117}$$

Here, $\bar{\pi}(b,\Psi)$ is assumed to be the *probability* of sampling two different alleles, rather than the average *number* of pairwise differences; these are approximately equal when π is small. This corresponds to an identity link function; one could alternatively use a two-alleles finite sites model link function of the form, $\pi/(1+2\pi)$. We experimented with this and found there was little difference between these link functions, so we opted for the simpler identity link function.

8 Simulations

8.1 Segment Simulations

9 Additional Theory

9.1 supp:weak-strong

10 Background Selection Models

The BGS models considered here both assume multiplicative fitness effects across sites. Thus, the total reduction due to segments across the genome is the product of the individual reductions,

$$B(x,\Psi) = \prod_{g}^{S} \exp\left(\int_0^1 b(\mu(s,k(g)),s,L_g,r_g,d(x,g))ds\right) \tag{118}$$

$$= \exp\left(\sum_{g}^{S}\int_0^1 b(\mu(s,k(g)),s,L_g,r_g,d(x,g))ds\right) \tag{119}$$

where, $b(\mu(s,k(g)),s,L_g,r_g,d(x,g))$ is the log reduction due to segment g, and there are S total segments. Here, $\mu(s,k(g))$ is the per-basepair per-genome rate that mutations with selection coefficient s enter segments with feature class $k(g)$, L_g and r_g are the length in basepairs and recombination rate per basepair of segment g, and the recombination distance between focal site x and segment g is $d(x,g)$. The functional form of b varies depending on whether the classic BGS model is used, or our form of Santiago and Caballero's (**Santiago2016-mu**) equations (see Section XXX).

Under both background selection models used in our likelihood, the diversity in window b is $\bar{\pi}(b,\Psi) = \bar{B}(b,\Psi)\pi_0$. Here, $\bar{B}(b|\Psi)$ is the predicted reduction in diversity due to BGS in window b, given background selection parameters Ψ. In practice, we pre-calculate $B(x|\Psi)$ at fixed sites x across the genome, and take the average of the fixed sites B values within widow b for the average $\bar{B}(b|\Psi)$.

10.1 Reductions Under the Classic BGS Model

The classic BGS model, which assumes deleterious mutations cannot fix, expresses the reduction for a focal site in the middle of a segment (**Hudson1995-xc**; **Hudson1994-oh**; **Nordborg1996-nq**); here we consider a focal site directly to an arbitrary side of a segment under a deleterious mutations, and extend the model to handle a focal segment arbitrarily far from the segment. To simplify notation, we will only consider a single feature class in this and the next section. Using our notation from equation (124), the classic mutation-selection-balance BGS model is,

$$b(\mu, s, L, r, 0) = -\int_0^L \frac{2\mu(l)}{s(1+(1-s)r(l)/s)^2} dl \tag{120}$$

(c.f. **Hudson1995-xc** equation 5), where $\mu(l)$ is the *haploid* per-basepair mutation rate at l and $r(l)$ is the recombination fraction between the focal site (at position 0) and basepair l. Assuming constant per-basepair recombination and mutation rates, $r(l) = r_{\text{BP}} l$ and $\mu(l) = \mu$, and

$$b(\mu, s, L, r_{\text{BP}}, 0) = -\frac{2\mu L}{s + (1-s) r_{\text{BP}} L} \tag{121}$$

(c.f. **Hudson1995-xc** equation 8, where the factor of two difference is due to the position of the focal site). Often per-region rates are defined $U = 2\mu L$ and $M = r_{\text{BP}} L$ are the segment-wide mutation diploid mutation rates (mutations per diploid segment per generation) and recombination length (Morgans).

For computational efficiency, we adjust this model by setting $r(l) = d + r_{\text{BP}} l$, where d is the recombination fraction between the focal site x and the start of segment g. This allows us to pre-compute the local effects of the segment, and substitute in b as the focal site changes. This integral substitution has a closed form,

$$b(\mu, s, L, r_{\text{BP}}, d) = \frac{-2\mu L}{(d(1-s)+s)(s + (d + r_{\text{BP}} L)(1-s))} \tag{122}$$

and terms can be collected in powers of d and pre-computed for all segments for the grid of μ and s. Our implementation pre-computes these segment components, and computes the final b value for an x and g by calculating the distance $d(x, g) = |M(x) - M(g)|$, where $M(x)$ is the cumulative recombination map length at position x in Morgans. Since $r(l)$ could vary by position, segments with differing recombination rates are split into segments of the same annotation class by their recombination rate. This has no effect on the calculation due to the multiplicative fitness, but ensures more accurate estimates of the reduction factor B.

10.2 Reduction Under the Modified Santiago and Caballero Model

Under Santiago and Caballero's model, the correct effective population size to gauge reduction in heterozygosity or coalescent times is given by equation XXX. This is because selective processes can lead to non-constant pairwise coalescent rates, which impact pairwise coalescent rates and thus the reduction factor. We experimented with the corrections described in Section XXX, and found (1) they had little effect on the accuracy of the B maps, and (2) they were incredibly costly to

calculate, even with the approximations described in equations XXX. Overall, we solve for the asymptotic $\widetilde{N}_{e,\infty}$ for each segment, and use that to calculate b'.

Note that the asymptotic model above determines the reduction in effective population size experienced by a focal site in the middle of a L-basepair segment, but B maps require the reduction b' experienced by a focal site x at an arbitrary recombination distance apart. For each segment, we pre-compute the solution to the system of two equations, giving us V and Q_∞^2, and then use these values in the equation (**??**),

$$b'(\mu, s, L, r, d) = -\frac{V}{2}\left(\frac{1}{(1-(1-d)Z}\right)^2 \tag{123}$$

where $Z = 1 - Us/(U - \widetilde{R})$ and $V = Us - 2\widetilde{R}s$. TODO check d.

10.3 Discretization of Parameters and Annotation Feature Classes

In practice, to fit such a model, we must discretize both $\mu(s, k(g))$ and s in Equation (124). We define μ_i to be the rate of mutations entering the population per basepair, per generation with selection coefficient s_i (i.e. the product of the mutation rate and the distribution of deleterious fitness effects). Then, the reduction at position x is,

$$B(x, \Psi) = \exp\left(\sum_g^S \sum_i b\left(\mu_{i,\,k(g)}, s_i, L_g, r_g, d(x, g)\right)\right). \tag{124}$$

Because it is computationally-intensive to numerically solve the system of two non-linear equations to calculate b at each segment, these are pre-computed for an m_s-element grid of selection coefficients and an m_μ-element grid of mutation rates–DFE products. The number of annotation features K is variable and set by the annotation data specified. There are two ways to parameterize the mutation rates and DFE for all annotation classes in background selection models. First, for the *free-mutation* parameterization, each annotation feature class has a free mutation rate parameter for each of these selection coefficients (as in equation (124)).

In the free-mutation parameterization, there are $1 + K \times m_s$ for π_0 and the free mutation rate parameters, where K is the number of annotation features. The MLE optimization for this approach is unconstrained, though π_0 is bounded to be within $10^{-5} \leq \pi_0 \leq 10^{-1}$ and mutation rates are bounded within $10^{-11} \leq \mu \leq 10^{-7}$. The BGS model requires a mutation rate for each selection coefficient in the grid, for each annotation class (here, as an example, CDS, UTRs, and phastcons), which are stored in the $m_s \times K$ matrix $\mathbf{M}_\mathrm{F}$,

$$\mathbf{M}_\mathrm{F} = \begin{bmatrix} \mu_{10^{-6},\mathrm{CDS}} & \mu_{10^{-6},\mathrm{UTR}} & \mu_{10^{-6},\mathrm{PC}} \\ \mu_{10^{-5},\mathrm{CDS}} & \mu_{10^{-5},\mathrm{UTR}} & \mu_{10^{-5},\mathrm{PC}} \\ \vdots & \vdots & \vdots \\ \mu_{10^{-1},\mathrm{CDS}} & \mu_{10^{-1},\mathrm{UTR}} & \mu_{10^{-1},\mathrm{PC}} \end{bmatrix} \tag{125}$$

That is, the distribution of fitness effect (DFE) is implied by the total mutation rate across selection coefficients for an annotation class. We can normalize by the total mutation rate estimate

to get the estimated DFE, giving us the DFE weight for annotation class k and selection coefficient with index i,

$$\widehat{w}_{i,k} = \frac{\widehat{\mu}_{i,k}}{\sum_i \widehat{\mu}_{i,k}}. \tag{126}$$

The second parametrization is the *simplex* parametrization, which has a single mutation rate across all features, so it is a DFE weight matrix times the mutation rate μ,

$$\mathbf{M}_{\mathrm{S}} = \mu \begin{bmatrix} w_{10^{-6},\mathrm{CDS}} & w_{10^{-6},\mathrm{UTR}} & w_{10^{-6},\mathrm{PC}} \\ w_{10^{-5},\mathrm{CDS}} & w_{10^{-5},\mathrm{UTR}} & w_{10^{-5},\mathrm{PC}} \\ \vdots & \vdots & \vdots \\ w_{10^{-1},\mathrm{CDS}} & w_{10^{-1},\mathrm{UTR}} & w_{10^{-1},\mathrm{PC}} \end{bmatrix} \tag{127}$$

which has $2 + K \times (m_s - 1)$ free parameters, since each column must sum to one. This imposes the constraint that $\sum_i w_{i,k} = 1$, requiring constrained MLE optimization.

11 Ratchet Rate Prediction

After rescaling each segment N_e by the local predicted reduction $\hat{B}(x)$, the ratchet rates and B' values are re-calculated.

The rescaled B' calculation outputs a $m_\mu \times m_s \times S$ multidimensional array $\mathbf{R}$ of ratchet rates per segment. These are rescaled by the segment lengths $\mathbf{l}$, giving the per-basepair mutation rate array. The per-basepair ratchet rate given maximum likelihood estimates of $\hat{\mu}$ and $\widehat{\mathbf{W}}$ is used to predicted the ratchet rate for each segment g. The DFE estimate for feature k is a column vector of $\widehat{\mathbf{W}}$, i.e. $\widehat{\mathbf{w}}_k$

$$\lambda_D(g) = \tag{128}$$

The predicted ratchet rate for segment g $\lambda_d(g)$ is based on these values.

11.1 Numeric Optimization

mu bounds

bounds and time, bounds and convergence

Same optima, but took longer