Characterising the detectable and invisible fractions of genomic loci under balancing selection

Débora Y. C. Brandt[1,2], Aida M. Andrés[1,2], Tim Connallon[4,*] and Max Reuter[1,3,*]

[1]Research Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, United Kingdom

[2]UCL Genetics Institute, University College London, Gower Street, London WC1E 6BT, United Kingdom

[3]Centre of Life Origins and Evolution, University College London, Gower Street, London WC1E 6BT, United Kingdom

[4]School of Biological Sciences, Monash University, 25 Rainforest Walk, Clayton, Victoria 3800, Australia

[*]Equal contribution

Abstract

Balancing selection refers to a suite of scenarios in which natural selection maintains genetic polymorphisms affecting fitness and its components. Empirical studies have documented a variety of specific cases of balancing selection, yet the general prevalence of balanced polymorphisms within genomes remains a topic of longstanding debate. Although genome-wide scans for signals of balancing selection suggest that it is rare, current methods are notoriously conservative and may identify only a small fraction of loci evolving under balancing selection. Consequently, the identified targets of balancing selection are unlikely to be representative of balancing selection in general. What, then, are the proportions of loci under balancing selection that are detectable versus invisible to genome scans? Here, we address this question using a combination of analytical modelling and population genetic simulations. Our results provide quantitative support for the intuition that a large fraction of loci under balancing selection will be invisible to standard tests for balancing selection, with the detectable loci representing a biased subset with large and symmetrical fitness effects on different fitness components. Quantifying the expected changes in power according to parameters that vary between species also shows that the detectable and invisible fractions of genomic loci under balancing selection are likely to differ substantially between organisms. For example, a *Drosophila melanogaster*-like ratio of mutation to recombination rates per nucleotide site (~0.1) reduces detection power by roughly a quarter of the power expected with a human-like ratio (~1). This observation, combined with other lines of evidence suggesting that balancing selection could be common in *D. melanogaster*, showcases how our appreciation of the prevalence of balancing selection is limited by our ability to detect its population genomic signals, especially in species where balancing selection might be common.

Introduction

Balancing selection occurs when natural selection actively favours the maintenance of genetic polymorphism. This can arise from several specific contexts of selection, including intrinsic fitness advantages of being heterozygous (*i.e.*, 'heterozygote advantage' or fitness 'overdominance'; (Fisher, 1922)), genetic trade-offs between different fitness contexts or environments (*e.g.*, 'antagonistic pleiotropy' between fitness components: (Curtsinger et al., 1994); 'sexually antagonistic' trade-offs between sexes: (Kidwell et al., 1977)); trade-offs between temporal or spatially varying environments (Haldane & Jayakar, 1963; Levene, 1953), or negative-frequency-dependent selection, where genotypes become more favourable when rare (Fitzpatrick et al., 2007; Wright, 1942).

Loci under balancing selection are expected to be more polymorphic than loci evolving neutrally or under directional selection (though exceptions can occur when selection is weak or allele frequency equilibria under balancing selection are close to zero or one, (Robertson, 1962; Williamson et al., 2004)). When equilibrium frequencies are intermediate (*e.g.*, very roughly between 0.2 and 0.8; (Hedrick, 1999)) and balancing selection is strong relative to genetic drift, the balanced polymorphisms should persist far longer than neutrally evolving polymorphisms (Nei & Roychoudhury, 1973), potentially over timescales spanning speciation events (Klein et al., 1998). Whatever their age, balanced polymorphisms should contribute disproportionately to standing genetic variation for fitness or fitness components (*e.g.*, for survival, fertility, or longevity), with even modest numbers of such loci contributing substantially to fitness variation within populations (Charlesworth, 2015; Connallon & Czuppon, 2026; Crow, 1952, 1987; Zajitschek & Connallon, 2018).

The importance of balancing selection has, nevertheless, been a subject of longstanding debate in evolutionary biology. As an early and vociferous proponent, Dobzhansky (1955) argued that selection should often preserve genetic variation due to the greater versatility of heterozygous, relative to homozygous individuals in perpetually changing environments. Yet, by the close of the 20th Century, high levels of molecular genetic diversity were reported by studies based first on allozymes (Lewontin & Hubby, 1966; Lewontin, 1974) and subsequently DNA sequences (Kreitman, 1983). The pervasiveness of these polymorphisms made Dobzhansky's 'balance hypothesis' an implausible general explanation for molecular genomic diversity. Instead, the emerging neutral theory (Kimura, 1968; King & Jukes, 1969; Kimura, 1983) provided a simpler and more powerful explanation for the ubiquity of molecular genetic

diversity, effectively sidelining balancing selection as irrelevant for all but a few exceptional loci within the genome (see p. 199 of Lewontin (1974); Hedrick (2012)).

More recent developments in evolutionary genetics have, however, prompted a resurgence in enthusiasm for balancing selection. Firstly, the coupling of field studies with genomics has identified exceptionally clear cases of balancing selection that maintains conspicuous phenotypic polymorphisms. Beyond classic examples like sickle cell anaemia (Allison, 1954), modern case studies include a pleiotropic trade-off between viability and mating success that maintains horn size polymorphism in Soay sheep (mapped to the *RXFP2* gene; Johnston et al. (2013)), a sexually antagonistic polymorphism in cuticular hydrocarbons of *Drosophila serrata* (in the *DsFAR2-B* gene; Rusuwa et al. (2022)), alleles affecting age of maturation in Atlantic salmon (the *VGLL3* gene; Barson et al. (2015)), temporally varying selection that maintains genetic polymorphism for insecticide resistance in *Drosophila melanogaster* (Karageorgi et al., 2025), and many others (reviewed in Bitarello *et al.* (2023); Ruzicka *et al.* (2025)). Second, the advent of high-throughput sequencing techniques and sophisticated analytical methods for detecting population genetic signatures of balancing selection has provided new means of identifying candidate loci under long-term balancing selection (reviewed in Bitarello et al., 2023; Fijarczyk & Babik, 2015). Applied to the human genome, such methods have provided strong evidence for long-term balancing selection in both African and European populations (Bitarello et al., 2018). Third, detailed lab studies of natural genetic variation in life-history traits suggest that *Drosophila* populations harbour higher levels of genetic variation than can be explained by models of mutation-selection balance (Charlesworth, 2015; Charlesworth & Hughes, 2000; Sharp & Agrawal, 2018). Balancing selection is a prime candidate to account for this apparent excess of genetic variation in life-history traits. If the same is true in other species, balancing selection may maintain substantial amounts of genetic and trait variation in natural populations.

While recent genome scans have identified putative balanced polymorphisms, it remains unclear how accurately they capture the true extent of balancing selection. The typically small fraction of loci with statistically significant signals of balancing selection (*e.g.*, <1% of the genome in humans (Bitarello et al., 2018)) might reflect either the rarity of balancing selection across the genome or methodological limitations to detecting it. Most notably, classical methods cannot detect balancing selection unless intermediate-frequency polymorphisms are maintained over exceptionally long timescales. Loci segregating for evolutionarily 'young' balanced polymorphisms cannot easily be identified through population genomic approaches alone (but see Isildak et al. (2021) for a potentially promising approach),

and old polymorphisms can likewise be missed due to limited data quality or other attributes that vary among species (*e.g.*, mutation and recombination rates; effective population sizes) and may influence the strength of balancing selection signals. The balanced polymorphisms that are detected are, therefore, likely to represent only a minor, potentially unrepresentative, subset of loci evolving under balancing selection. In fact, a recent genome-wide analysis suggests there are hundreds of non-synonymous polymorphisms shared among human populations due to recent balancing selection, even though the method can only detect an aggregate signal across many loci rather than identify individual targets of balancing selection (Soni et al., 2022). While it remains an open question what proportion of these loci ultimately results in long-term stable balanced polymorphisms that are individually identifiable, this analysis suggests that at least recent episodes of balancing selection are widespread in the human genome.

Here, we aim to put the debate over the prevalence of balancing selection on a sounder conceptual and quantitative footing by evaluating factors affecting opportunities for long-term balancing selection and its detectability. It is obviously impossible to make general predictions about the proportion of a genome that is a potential target of balancing selection. Yet it is possible to assess—under a given model of balancing selection—how the population genetic parameters of genetic drift and selection will affect: *i*) the long-term maintenance of polymorphisms, and *ii*) the proportion of loci that develop signatures of balancing selection that are clear enough to be detectable using contemporary population genomic methods.

We address these questions while considering three mechanisms of balancing selection: overdominance, antagonistic pleiotropy and sexually antagonistic selection, each of which have been empirically documented in natural populations (e.g. Allison, 1954; Johnston et al., 2013; Rusuwa et al., 2022). These mechanisms represent a subset of balancing selection mechanisms (e.g. Glémin, 2021; Prout, 2000; Ruzicka et al., 2025), yet their evolutionary dynamics are representative of a broader array of balancing selection scenarios, which allows our predictions to be generalized. We begin by briefly reviewing predictions from population genetics theory, highlighting constraints associated with maintaining long-term stable polymorphisms under each mechanism of balancing selection. We then present simulations under each mechanism that explore the genomic signals associated with long-term balanced polymorphisms and assess how these signals translate into statistical power to infer the presence of balancing selection at these loci. Our analyses reveal how the maintenance of balanced polymorphisms and their detectability depend on the properties of the genetic polymorphism itself, as well as the genetic and demographic features of the species. They also

provide a quantitative appreciation for the extent to which loci subject to balancing selection are detectable and suggest the biases in the specific types of balanced polymorphisms that are most and least likely to be identified.

Results

Long-term maintenance of balanced polymorphisms: Theoretical considerations

We start by reviewing theoretical predictions regarding the conditions under which genetic polymorphisms are stably maintained due to overdominant selection (OD), antagonistic pleiotropy (AP) or sexually antagonistic selection (SA). Our models focus on bi-allelic, diploid loci, and we assume that individuals mate at random and selection at each locus remains constant over time. Violations of these assumptions tend to restrict conditions for balancing selection (reviewed in Ruzicka et al. (2025), but see Wittmann et al. (2023) for signals of long-term balancing selection under temporally fluctuating selection). Full details of the models are presented in the Supplementary Material.

Long-term maintenance of a polymorphism under these models of selection requires two conditions. First, the combination of allelic fitness effects must actually generate balancing selection. Second, the resulting selection must be stronger than genetic drift for the polymorphism to be stably maintained over time. We deal with each condition in turn.

As far as the general criteria for balancing selection are concerned, OD is the most permissive scenario. In outbred populations, heterozygotes having higher fitness than both homozygotes (OD) invariably leads to balancing selection (Fisher, 1922). In contrast, AP and SA lead to balancing selection under only a subset of parameter conditions (Fig. 1). AP and SA each involve trade-offs between the fitness effects of alleles in different contexts of selection (between fitness components and between sexes, respectively). Considering a locus with alleles A_1 and A_2, s_1 represents the homozygous cost associated with the A_1 allele, s_2 is the homozygous cost associated with the A_2 allele, and dominance (h) defines the degree to which these costs manifest in heterozygotes (Fig. 1A). For simplicity, we focus on cases in which the dominance coefficient ranges from $h = 0.5$, implying co-dominance of the alleles, to $h = 0$, implying complete recessivity of these costs (for generalizations beyond these conditions, see Kidwell et al. (1977); Curtsinger et al. (1994)). Note that dominance values within the range $0 < h < 0.5$ correspond to "dominance reversals" (Connallon & Chenoweth, 2019; Grieshop et

al., 2024; Reid, 2022), where each allele is (partially) recessive in contexts where it is harmful and (partially) dominant in contexts where it is favourable.

Balancing selection arises consistently in the SA and AP scenarios when the homozygous costs of each allele are symmetric in magnitude (s_1 is roughly equal to s_2), and the criteria for balancing selection expand with both the strength of selection and the strength of the dominance reversal (i.e., with s_1 and s_2 increasing and h decreasing; Fig. 1B). When selection is weak and there is no dominance reversal, only a small proportion of the parameter space will generate balancing selection. In this case, the proportion of random parameter combinations that leads to balancing selection is roughly $f_{bal} \approx \frac{2}{3}\bar{s}$ under AP and $f_{bal} \approx \frac{4}{3}\bar{s}$ under SA, where $\underline{s}$ represents the average fitness effect ($\bar{s} = \frac{s_{\max}}{2}$ for parameters drawn from uniformly from the parameter space, with $s_{\max}$ representing the maximum selection coefficient; $s_{\max} \ll 1$ when selection is consistently weak). With dominance reversals ($h < 0.5$) and weak selection, the proportions expand to approximately $f_{bal} \approx \frac{1-2h}{1-h}$ for both AP and SA, which corresponds to $f_{bal} \approx \frac{2}{3}$ when $h = 0.25$ (as in Fig. 1B) and $f_{bal} = 1$ for a complete dominance reversal ($h = 0$).

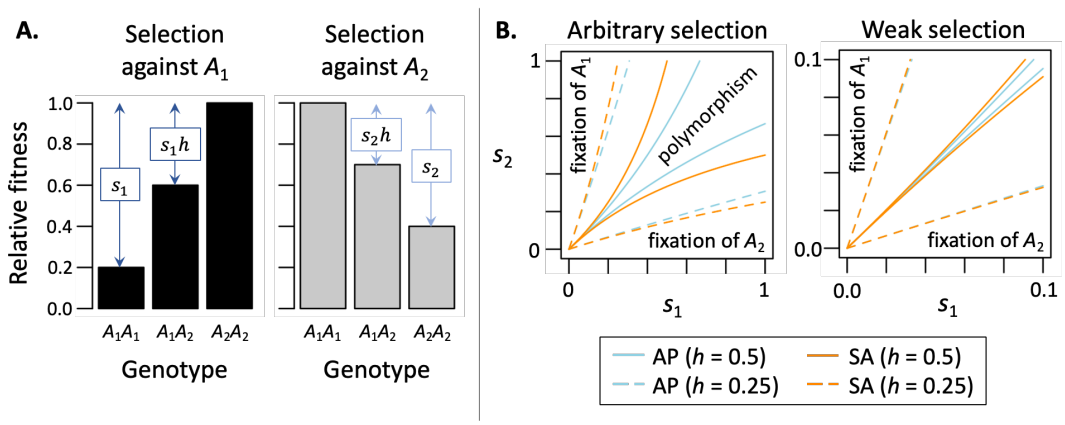

Figure 1. Conditions for balancing selection under antagonistic pleiotropy (AP) and sexually antagonistic selection (SA). **Panel A:** Visual summary of the AP and SA model parameters, with the example showing $s_1 = 0.8$, $s_2 = 0.6$, and $h = 0.5$. The two bar charts show the relative fitness of each of the three genotypes of a locus across two contexts of selection (i.e., different fitness components or sexes), where the A_1 allele is selectively disfavoured in one context and the A_2 allele is disfavoured in the other. **Panel B:** Criteria for balancing selection as a function of selection (s_1 and s_2) and dominance (h) parameters. The regions between each pair of curves (based on

colour and line type) show conditions where selection favours the maintenance of polymorphism (i.e., where there is balancing selection). Predictions are shown when there is no dominance between alleles (i.e., alleles are co-dominant: h = 0.5), and when there is a beneficial reversal of dominance (h = 0.25), where negative allelic effects are partially recessive to beneficial effects. Panel B is modified from Connallon and Chenoweth (2019), with curves based on standard modelling results (see Kidwell et al. 1977; Curtsinger et al. 1994).

The second condition for long-term stability of a balanced polymorphism is that selection towards a polymorphic equilibrium (which we denote as $\hat{p}$) must be strong enough to prevent fixation of one or the other allele by genetic drift. The required strength of selection depends on the effective size of the population (N_e) and the net strength of selection to the equilibrium (α), which differs between the mechanisms of balancing selection ($\alpha = s_1 + s_2$ under the OD model, where s_1 and s_2 represent the fitness reductions of A_1 and A_2 homozygotes relative to the heterozygotes; $\alpha = (s_1 + s_2)(1 - 2h) + 2s_1 s_2 h^2$ and $\alpha = \frac{1}{2}(s_1 + s_2)(1 - 2h) + 2s_1 s_2 h^2$ under the AP and SA models, respectively, with the SA and AP parameters, s_1, s_2, and h, as depicted in Fig. 1A; see Supplementary Text for derivations of the equations for α).

The stationary distribution of the allele frequencies for a locus under balancing selection and drift (*i.e.*, the predicted long-term probability distribution of allele frequency states) provides an intuitive measure of balancing selection's ability to maintain polymorphism in the presence of genetic drift (Rice, 2004; Robertson, 1962). Balancing selection reliably maintains polymorphism when the allele frequency probability densities are much greater near the deterministic equilibrium ($\hat{p}$) than near the boundaries (*i.e.*, close to zero and one). Heuristically, balancing selection can be regarded as effectively strong in cases where the probability density around $\hat{p}$ is at least ten times greater than the density near the boundaries. Under this definition, balancing selection is effectively strong when $\alpha > a_{\min}$, where:

$$a_{\min} = \begin{cases} \dfrac{\log(20N_e\hat{p})}{2N_e\hat{p}^2} & \text{for } \hat{p} \leq 0.5 \\ \dfrac{\log\big(20N_e(1-\hat{p})\big)}{2N_e(1-\hat{p})^2} & \text{for } \hat{p} \geq 0.5 \end{cases} \tag{1}$$

Fig. 2A illustrates how the effective population size and equilibrium allele frequencies under balancing selection (N_e and $\hat{p}$) interact to determine criteria for effectively strong balancing selection (where $\alpha > a_{\min}$). As can be seen in the figure, these conditions are most permissive

in large populations with intermediate equilibria (note that $a_{\min}$ is always minimized at $\hat{p} = 0.5$).

Because α is defined differently for each model of selection (see above), thresholds for effectively strong balancing selection correspond to different absolute strengths of selection under the AP, SA, and OD scenarios. For example, the minimum values of s_1 and s_2 that are required for $\alpha > a_{\min}$ (shown in Fig. 2B) are largest for the SA model, smallest for the OD model, and intermediate for the AP model (SA and AP cases are equal when $h = 0.5$, and AP and OD cases are equal when $h = 0$; see Fig. 2B).

A reduction of N_e amplifies drift and restricts the parameter space of effectively strong balancing selection (Fig. 2C), particularly so when dominance reversals are weak or absent. For example, when $h = 0.5$ and $N_e = 2{,}000$, SA and AP alleles cannot be reliably maintained polymorphic even by moderately strong selection (*e.g.*, $s_1 \sim s_2 \sim 0.05$, corresponding to a 5% difference between the homozygote genotypes in each fitness component). Larger N_e and/or dominance reversals ($h < 0.5$) make long-term maintenance of polymorphism much more likely (Fig. 2C). Population genetic simulations (see Methods for simulation details) confirm these general predictions (Fig. S1A,C). They further illustrate how the notion of maintaining long-term stable polymorphisms depends on the timescale of the observation: the range of parameter values that maintains polymorphism for $16N_e$ generations is significantly narrower than the range that maintains polymorphism for $8N_e$ generations (Fig. S1C,D).

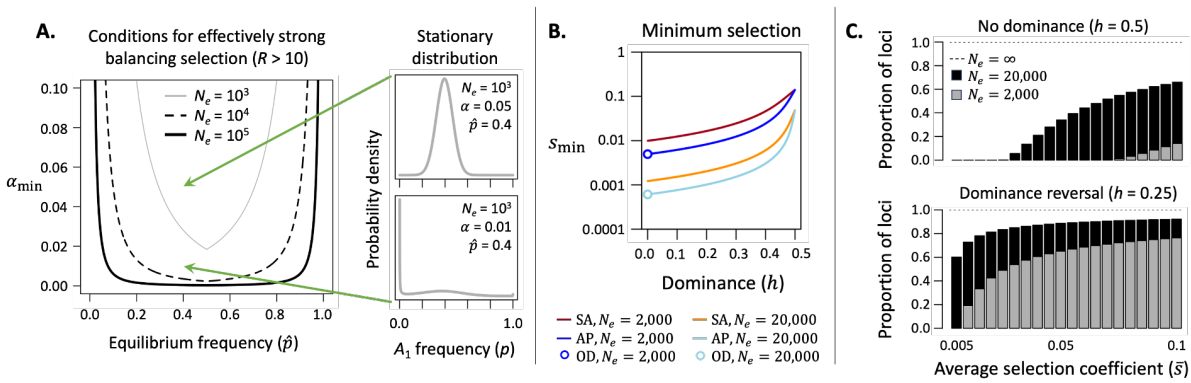

Figure 2. Effectively strong balancing selection. **Panel A** illustrates the thresholds for effectively strong balancing selection relative to genetic drift ($a_{\min}$; see eq. (1) and the surrounding text) for three population sizes (N_e) and a spectrum of allele frequency equilibrium conditions under balancing selection ($0 < \hat{p} < 1$). Regions above each curve denote combinations of α and $\hat{p}$ that generate effectively strong

balancing selection (i.e., where the probability density for the stationary distribution is at least a 10-fold higher near the polymorphic equilibrium, $\hat{p}$, than near the allele frequency boundaries of zero and one, which correspond to fixation). Two special cases of the stationary distribution are shown: one where selection is ineffective ($\alpha < a_{\min}$, and density is concentrated near *zero*) and one where selection is effective ($\alpha > a_{\min}$, and density is concentrated near the equilibrium, $\hat{p} = 0.4$). **Panel B** shows the minimum selection coefficients for each model that allow for effective balancing selection (the minimum is defined at $\hat{p} = 0.5$, where conditions for effective balancing selection are most permissive). **Panel C** shows the proportion of loci under AP balancing selection that meet the criteria for effectively strong balancing selection. Results are based on randomly drawing 10^7 sets of selection coefficients (s_1 and s_2) from independent uniform distributions between 0 and $s_{\max}$, where $\bar{s} = \frac{1}{2} s_{\max}$ is the average selection coefficient. Bar heights were calculated as the number of parameter sets that met the criterion for effectively strong balancing selection ($0 < \hat{p} < 1$ and $\alpha > a_{\min}$), divided by the number of parameter sets that generated balancing selection ($0 < \hat{p} < 1$).

Empirical signatures of long-term balancing selection

The theory presented above summarizes factors that promote or constrain the maintenance of polymorphism at individual sites subject to balancing selection. Assessing the detectability of balancing selection requires additional considerations, because genomic scans for balancing selection do not focus on the balanced site itself but on signals of elevated genetic diversity at sites physically linked to the selected site. These signatures build up in the presence of long-term balancing selection, where the balanced polymorphism segregates for much longer than expected under neutrality (*i.e.*, the time to the most recent common ancestor, TMRCA, at the selected site is greater than the neutral expectation of $4N_e$ generations). Genomic regions with an old TMRCA owing to balancing selection are expected to deviate from the genomic background in two ways. First, they are expected to show an elevated density of polymorphisms (and a concomitant deficit of substitutions relative to closely related outgroups) due to the long timespans over which mutations can accumulate since the TMRCA in the region and their lower probability of going to fixation due to linkage with the balanced site (DeGiorgio et al., 2014). Second, they are expected to show an excess of intermediate-frequency alleles

when compared with the neutral expectation, as linked neutral variants tend to segregate at frequencies similar to those of the balanced polymorphism (see e.g., Fig. 1 in Bitarello et al. (2018)). These properties will be reflected in the local site-frequency spectrum (SFS) and provide the signal exploited by current population genomic methods for detecting long-term balancing selection (e.g. Bitarello et al., 2018; Cheng & DeGiorgio, 2020).

In the following section, we will explore how the presence and intensity of the signal, and thus our statistical power to identify regions under balancing selection, depends on the properties of the balanced polymorphism itself (*i.e.*, the nature and strength of balancing selection, dominance, and the resulting equilibrium frequency of the balanced polymorphism), and factors that affect population genomic data, including the effective population size, and mutation and recombination rates, which vary among species and genomic regions.

We can get a preliminary sense of some of the key factors affecting neutral diversity near a long-term balanced polymorphism by quantifying the expected spike in heterozygosity near a balanced site under strong selection, which is analytically tractable. Consider a balanced polymorphism with alleles A_1 and A_2 maintained stably at deterministic equilibrium frequencies, $\hat{p}$ and $1-\hat{p}$, respectively. Building on previous work (see p. 103 in Rice (2004)), one can derive the expected width of the sequence surrounding the balanced site in which neutral polymorphisms will have an expected level of heterozygosity that is at least K times greater than the expected heterozygosity at unlinked neutral sites elsewhere in the genome ($4N_e\mu$, where μ is mutation rate per nucleotide site). The width of this region, in kilobases (kb), is:

$$L_K \approx \frac{2\hat{p}(1-\hat{p})}{10^3(K-1)N_e r} \tag{2}$$

where r represents the recombination rate per nucleotide site (see Appendix 3). Inspection of eq. (2) shows that the width of elevated diversity surrounding the balanced polymorphism increases with the equilibrium heterozygosity of the selected site (with $2\hat{p}(1-\hat{p})$), which is maximized when balanced alleles are maintained at equal frequencies ($\hat{p}=0.5$), and it decreases with the population-scaled recombination rate per nucleotide site ($N_e r$). Thus, we expect empirical signals of long-term balancing selection to be strongest when $\hat{p}$ is intermediate and $N_e r$ is small, owing to small N_e and/or small r.

Power to detect balancing selection in simulated data

While these predictions lend intuition, they ignore the reality that allele frequencies at the selected site fluctuate over time (*e.g.,* due to drift), and overlook other empirically distinctive

features of the SFS that are relevant to balancing selection scans, including the actual number of segregating linked variants and the shape of the SFS around a balanced site relative to the genome as a whole. Allele frequency fluctuations at the balanced polymorphic site are expected to reduce linked diversity relative to the more idealized case (as in eq. (2)) where the alleles are maintained at invariant equilibrium frequencies (for allele frequency fluctuations caused by temporally varying selection, see Wittmann et al. (2023)). Drift-mediated allele frequency fluctuations at the selected site should be minimal provided $N_e\alpha$ is large and the equilibrium ($\hat{p}$) is intermediate, while fluctuations increase as $N_e\alpha$ and $\hat{p}(1-\hat{p})$ decline (Charlesworth & Charlesworth, 2010, pp. 354–355; Connallon & Czuppon, 2026).

To provide an assessment of statistical power for detecting balancing selection candidates, we turned to population genetic simulations. We ran extensive simulations (see Methods for details) with population-scaled values of $N_e\mu$ and N_er chosen to roughly match those of humans and fruit flies—intensively studied species that may be representative of vertebrates and insects. We note that these simulations are not intended to provide specific power estimates for these species, since our models do not match their demographic histories (for example, the long-term N_e of human populations is smaller than 20,000 and their detailed demographic history generally increases power compared with the stable effective population size simulated here see Discussion). Rather, they are presented as toy examples that allow us to evaluate effects of relevant genomic parameters that vary among species. We first assessed power for a human-like 'reference' set of parameters (N_e = 20,000, $\mu = r = 10^{-8}$, age of polymorphism $T = 8N_e$ =160,000 generations), which we will later use as a point of comparison with simulations using different population parameters. We assessed the power of the NCD method (Bitarello et al., 2018) to identify balancing selection in population genetic simulations of 10Kb chromosomes with a centrally located site under balancing selection and neutral variation in the flanking regions (see Methods). The results obtained for NCD should be representative of those obtained with other contemporary methods for detecting long-term balancing selection, which use similar approaches; additional results using the BalLeRMix method (Cheng & DeGiorgio, 2020) are presented in the Supplementary Material. NCD tests for an enrichment of polymorphism in the SFS by calculating, across all SNPs in a window, the average deviation of allele frequency from an intermediate 'target frequency'. We used the equilibrium frequency of the balanced site as the target frequency for NCD, and our analyses therefore represent an idealised best-case scenario for power, where the true equilibrium frequency is known (Bitarello et al., 2018). For each set of parameters, we generated 100 replicate simulations in which the balanced polymorphism segregated until the end of the

355 simulation ($8N_e$ or $16N_e$ generations, thus conditioning on the maintenance of polymorphism;
356 for results on the probability of maintenance over these time durations, see Fig. S1-S4). From
357 these, we estimated power as the proportion of simulations that resulted in a statistically
358 significant signal of balancing selection at $P < 0.01$, measured by NCD.

359 We evaluated the effects of our simulation parameters on power through multiple
360 logistic regressions (Generalised Linear Models (GLM) with binomial error) with counts of
361 successful and unsuccessful detection among the 100 trials of our simulations as dependent
362 variables. The GLMs model power on the log-odds scale and allow us to estimate and test the
363 effects of the simulation parameters while taking into account the fact that power is bounded
364 between 0 and 1 as well as the binomial error of successful/unsuccessful detection of balancing
365 selection (see Methods for more details).

366 In our simulations, the power to detect balancing selection is relatively low overall for
367 our set of parameters, regardless of the exact method used for detection (i.e., NCD in Fig. 3
368 versus BalLeRMix in Fig. S8-10). However, power varied considerably across parameter
369 conditions in ways that broadly align with theoretical expectations. Among our human-like
370 reference simulations, the variable that most strongly predicts power is the equilibrium
371 frequency of the balanced polymorphism ($\hat{p}$), with the likelihood of detecting balancing
372 selection increasing as the equilibrium approaches an intermediate frequency (i.e., $\hat{p} \rightarrow 0.5$;
373 Fig. 3A). Across these reference simulations, the average power is 0.23 at $\hat{p} = 0.5$, while power
374 drops to 0.02 with $\hat{p} = 0.07$. Fitting a binomial GLM with equilibrium frequency ($\hat{p}$), selection
375 strength (α), selection model and their interactions as predictors of detection power, we found
376 that the equilibrium frequency explains more than 60% of the variance in power across our
377 simulations (Table 1). More specifically, the odds of successfully detecting balancing selection
378 are estimated to increase by more than 30% with every 10% increase in the equilibrium
379 frequency of the minor allele at the selected site, up to $\hat{p} = 0.5$ ($b_{eq.\ freq.} = 2.737$ on log-odds
380 scale, estimated for OD).

381 Power also generally increases with the strength of selection (Table 1), but this effect
382 depends on interactions with other parameters. First, the positive effect of selection strength on
383 power is pronounced at intermediate equilibrium frequencies but declines and eventually
384 vanishes when equilibrium frequencies of the minor allele at the selected site are lower (Table
385 1, Fig. 3B). Second, this interaction effect differs between the three models of selection and is
386 strongest under SA. Under OD and AP, increasing selection strength by 0.1 increases the odds
387 of detecting balancing selection by 5.7% and 1.6% for equilibrium frequencies of $\hat{p} = 0.5$ and

$\hat{p}$ = 0.25, respectively. For SA, in contrast, the same change in selection strength increases the odds of detection by 15.7% for an equilibrium frequency of $\hat{p}$ = 0.5 but decreases the odds by 40% for an equilibrium of $\hat{p}$ = 0.25 (see Fig. 3B for simulation data and Fig. 4 for model predictions). Beyond these interactive effects, however, the model of selection does not impact the overall power of detecting balancing selection (Table 1, Fig. 3C).

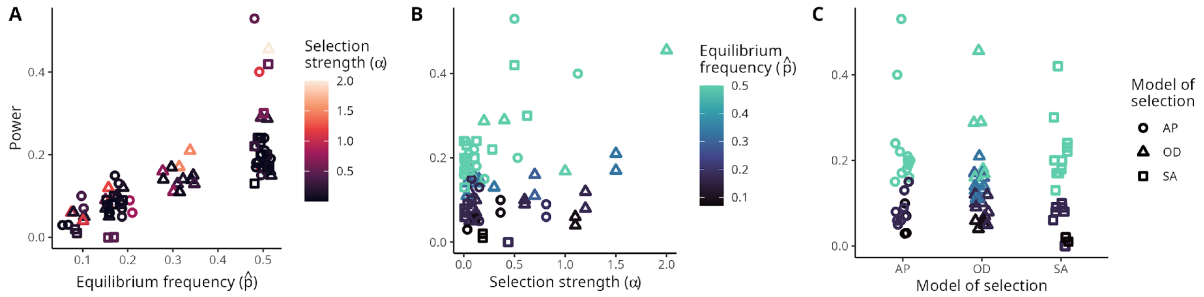

Figure 3. Power as a function of equilibrium frequency ($\hat{p}$, panel A), selection strength (α, panel B), and model of selection (panel C) for the reference simulations. On panels A and C points were horizontally jittered for visualization purposes. The three panels show the same data points, with power on the vertical axis plotted against different horizontal axes of variation.

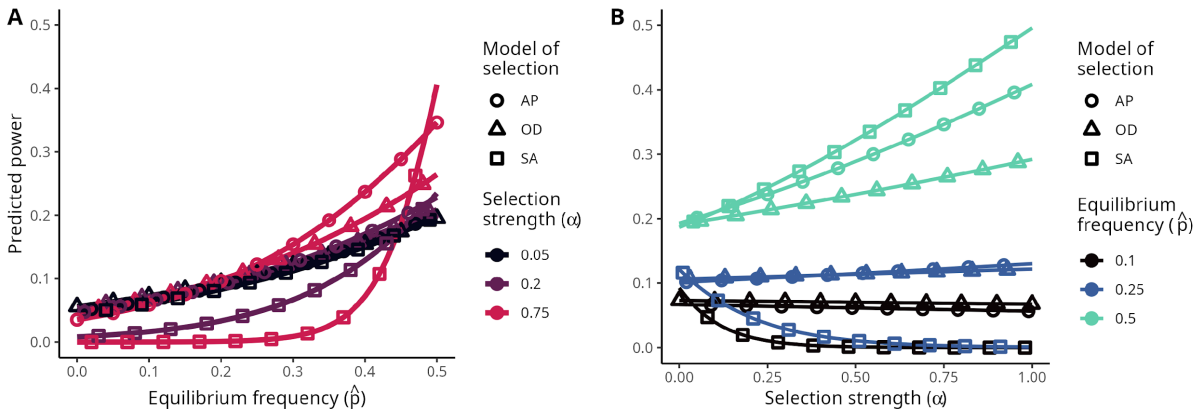

Figure 4. Power predicted by the Generalised Linear Model fitted to the data for standard simulations as a function of equilibrium frequency ($\hat{p}$, panel A) and selection strength (α, panel B). Results are shown for different parameter combinations as indicated by the right-hand legend of each panel.

Table 1. Binomial GLM of numbers of instances of successful/unsuccessful detections of balancing selection in the reference set of simulations, as a function of equilibrium frequency ($\hat{p}$), selection strength (α), selection model and their interactions. Only terms with P-value <0.05 are shown.			
Predictor	Deviance	% deviance explained	P-value
$\hat{p}$	362.96	62.28%	< 0.001
α	47.66	8.2%	< 0.001
$\alpha * \hat{p}$	13.64	2.3%	< 0.001
$\hat{p}$ * selection model	13.68	2.4%	0.001
$\alpha * \hat{p}$ * selection model	33.38	5.7%	< 0.001

As current methods for inferring targets of long-term balancing selection rely largely on a signal of excess polymorphism and intermediate frequencies in the SFS, power should increase in conditions that generate a larger and wider signature of excess polymorphism and decrease under those that restrict the signal. To explore these effects, we ran additional simulations under OD while varying the recombination and mutation rates, the effective population size, and the age of the balanced polymorphism. These simulations show that power consistently decreases at higher recombination rates (Table 2), particularly so in cases where the equilibrium frequency is intermediate (Fig. 5A; Table 2). The model predicts that a ten-fold increase in recombination rate decreases the odds of detection by roughly ten-fold when the equilibrium is intermediate ($\hat{p} = 0.5$), but only by 4.5-fold for an equilibrium of $\hat{p} = 0.25$ (Table 2, b values).

Conversely, power increases with a higher mutation rate per nucleotide site (Figure 5B). The model predicts that the odds of successfully detecting balancing selection increase by about 150% for each ten-fold increase in mutation rate (Table 2, b value). Note that while there is no interaction of mutation rate with the allele frequency equilibrium (Table 2), there is a nominally significant triple interaction with the equilibrium frequency and strength of selection, though this explains a negligible proportion of variation in power (Table 2).

The intensity of the signal of balancing selection also depends on the timespan over which a balanced polymorphism has been maintained, with the age of the balanced polymorphism in our simulation data having a strong positive effect on our power to detect it (Fig. 5C; Table 2). In our analyses, the allelic age also interacts significantly with the equilibrium frequency in determining power (Table 2). Accordingly, for every additional 0.5 N_e generations of allelic age in our model, the odds of successfully detecting balancing selection increase by 9.5% for a polymorphism with an equilibrium frequency of 0.5 and by 5.5% with an equilibrium frequency of 0.25.

Finally, increasing the effective population size generally reduces the detectability of balanced polymorphisms, though effects of N_e are complex (Fig. 5D). The binomial GLM reveals a significant negative effect of N_e on power (Table 2), with a significant interaction with the equilibrium frequency and with the interaction of equilibrium frequency with selection strength (Table 2). Thus, the GLM predicts that among loci with a minor allele equilibrium frequency close to our lower bound of $\hat{p} = 0.1$, and under weak selection ($\alpha = 0.1$), a ten-fold increase in N_e lowers the odds of successfully detecting balancing selection by over 30%. In contrast, for loci with an intermediate equilibrium frequency of $\hat{p} = 0.5$ and under strong selection ($\alpha = 1.5$), the GLM predicts that the same increase in N_e would reduce the odds of detection by only 10%.

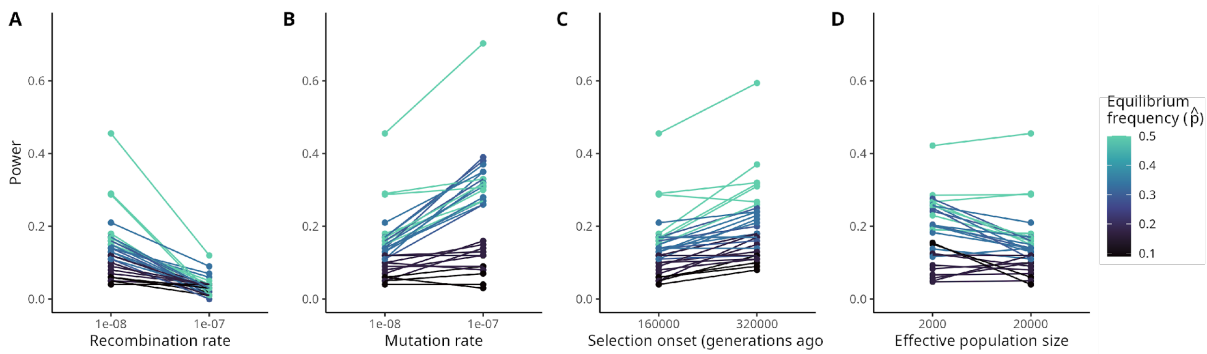

Figure 5. Power to detect balancing selection in our reference set of simulations (points on the left of panels A-C and on the right of panel D) compared to simulations with increased recombination rate, mutation rate, age of the selected allele and reduced effective population size. Balancing selection was simulated under overdominance, and parameters in the reference set of simulations are N_e = 20,000, recombination and mutation rates of 10^{-8} per site per generation, selection onset time 160,000 generations ago.

451

Table 2. Binomial GLM of numbers of instances of successful/unsuccessful detections of balancing selection (overdominance model) as a function of predictors and interactions. Only terms including the variable of interest and with P-value <0.05 are shown.					
Predictors included in the model	Predictors discussed in main text	Deviance	% deviance explained	P-value	b (odds ratio)
equilibrium frequency ($\hat{p}$), selection strength (α), $\log_{10}$ of **recombination rate (r)**	r	249.89	55.3%	<0.001	-0.841 (0.43)
	r * $\hat{p}$	11.12	2.5%	0.001	-2.766 (0.06)
equilibrium frequency ($\hat{p}$), selection strength (α), $\log_{10}$ of **mutation rate (m)**	m	112.86	18.9%	<0.001	0.879 (2.41)
	m * $\hat{p}$ * α	3.98	0.7%	0.046	1.776
equilibrium frequency ($\hat{p}$), selection strength (α), **allelic age (t)**	t	557.1	51.2%	<0.001	0.002 (1.002)
	t * $\hat{p}$	69.17	6.4%	<0.001	0.015 (1.015)
	t * α	12.50	1.1%	<0.001	-0.002
equilibrium frequency ($\hat{p}$), selection strength (α), **effective population size (N_e)** in thousands of generations	N_e	82.02	23.2%	<0.001	-0.036 (0.96)
	N_e * $\hat{p}$	17.92	5.1%	<0.001	0.033 (1.03)
	N_e * $\hat{p}$ * α	6.51	1.8%	0.011	0.130 (1.14)

452

453

Discussion

We assessed the propensity of balancing selection to maintain long-term stable polymorphism and the power of current inference methods to successfully detect such loci. The main message from our analyses is that detecting balanced sites in population genomic data can be difficult, at least in the populations of constant size that we simulate here. Even where balancing selection is strong relative to drift and has maintained allelic variation at frequencies closely clustered around an equilibrium frequency of 0.5 for time spans that vastly exceed neutral coalescent times (such as $16N_e$ generations), state-of-the-art methods of inference still only rarely exceed a power of 50%. For example, a mutation under OD with relatively strong and symmetrical selection ($s_1 = s_2 = 0.1$, resulting in $\hat{p} = 0.5$) that is $8N_e$ generations old is detected as a balanced locus with a probability of only 28% in our reference simulations. These selection coefficients are of the same order of magnitude as the ones estimated for the remarkable example of net overdominance maintaining horn size polymorphism in Soay sheep (Johnston et al. (2013), see their Table S3) . Our results thus show that in our models, 72% of loci under similar selection strengths and equilibrium frequencies would be missed. This finding contrasts with the notion that inference power with current approaches is good (Bitarello et al., 2023), a perception that is probably due to the empirical focus on human populations, which have particular demographic histories that are generally expected to boost the power of existing methods (see below).

While power in our simulations is moderate in the best of cases, it further decreases, often substantially, under less ideal and perhaps more realistic conditions. The selection parameters have particularly pronounced effects on the long-term maintenance and eventual detectability of a balanced locus. Selection and dominance coefficients determine whether a polymorphism will be subject to balancing selection at all (as in the AP and SA cases), the equilibrium frequencies of balanced polymorphisms, and the efficacy with which selection maintains the alleles near equilibrium. Each of these factors influences the long-term stability of a balanced polymorphism. But even when a polymorphism is long-term stable, balancing selection resulting in allele frequency equilibria close to zero or one are exceedingly unlikely to be identified, not only because they are more susceptible to becoming lost by drift, but also because even those polymorphisms that persist produce weak signals of balancing selection that are very hard to identify.

Model of selection mainly affects maintenance

The models of selection differ very strongly in the strength of selection they generate and hence their propensity to maintain polymorphism in the first place. With OD the net strength of selection is a simple linear function of its underlying selection coefficients ($\alpha = s_1 + s_2$), leading to effectively strong net selection in cases where population-scaled selection parameters are large (i.e., $N_e s_1 \gg 1$ and $N_e s_2 \gg 1$). In contrast, for scenarios of AP and SA without dominance reversals, the net strength of balancing selection (when it occurs) depends on the product of the selection coefficients (i.e., $\alpha = \frac{1}{2} s_1 s_2$ when h = 0.5), so that the same values of s_1 and s_2 lead to much weaker balancing selection under AP and SA than under OD (see Appendix 2 for derivations of equations for α). Assuming the distribution of fitness effects is similar under the three mechanisms, then those subject to antagonistic selection should be less likely to remain polymorphic over extended periods of time and, where they do, be harder to detect by genome scans for balancing selection. While dominance reversals dampen these differences between the antagonistic selection scenarios and OD, balancing selection in the SA and AP scenarios are expected to remain weaker and less permissive than the OD case unless dominance reversals are complete.

While the specific model of balancing selection has a strong effect on the propensity to maintain balanced polymorphisms, it did not have an overall effect on detection power. However, our analyses revealed more subtle interactions between the model of selection, the net strength of selection (α), and the equilibrium frequency ($\hat{p}$). Specifically, at low equilibrium frequencies, power decreases with increasing strength of selection under SA, while selection strength has little or no effect on power under OD and AP. This effect is surprising, given that α is intended to capture the strength of selection in a way that is independent of the mechanism generating it. We note, however, that the expressions for α are approximations based on the assumption of effectively weak selection (see Appendix 1). It is therefore possible that the interaction effects we detected here were due to higher-order terms that are not included in the calculation of α.

Selection onset and effective population size

Our simulations also demonstrate that the probability of inferring balancing selection at a locus increases with the age of the polymorphism, as signatures of linked variation build up over

517 time. Importantly, the timescale of this process increases with the effective size of the
518 population. For the signature of selection in linked variation to be discernible, the coalescent
519 time at the balanced site needs to substantially exceed the average neutral coalescent time of
520 $4N_e$ generations. This means that when N_e is larger, balanced polymorphisms need to be older,
521 in generations, to be detectable. Likewise, for a given age in generations, polymorphisms are
522 harder to detect in species with high N_e than in those with low N_e.

523 But the timescales over which signatures of selection accumulate, and how they differ
524 between species, also matter for the presence of balancing selection more generally. Thus, the
525 time for signatures to build up can be considerable in geological time. In humans, for example,
526 an age of $8N_e$ generations corresponds to roughly 2My (based on a historical N_e of 10^4 and a
527 25-year generation time), or roughly a quarter of the time since the human–chimpanzee split.
528 For a polymorphism to be present and detectable after this time, sizable fitness effects need to
529 persist for this duration. This could substantially limit the prospects for the presence of long-
530 term balancing selection in species with large N_e, since stability of the polymorphism requires
531 a relative constancy of the environment over a long time. However, this effect is attenuated by
532 the fact that generation times tend to correlate negatively with N_e. An analysis of the
533 relationship between N_e and generation time (g) (Chao & Carr, 1993) suggests that, very
534 roughly, $g \propto N_e^{-0.35}$, in which case the geological timescale (in years) that is required to
535 generate a genomic signal of long-term balancing selection is on the order of $N_e g \propto N_e^{0.65}$.
536 Thus, despite the inverse relationship between N_e and g, the timescale for generating a signal
537 of balancing selection should generally be higher in lineages with larger N_e and shorter g (such
538 as insects). As a consequence, for a given geological age of a polymorphism, the signals of
539 balancing selection are less likely to be visible in large-N_e lineages because the criteria for
540 "long-term balancing selection" (as defined in Bitarello et al., 2023) are more restrictive.

541 Interestingly, *Drosophila melanogaster* may be an exception to the broader pattern, in
542 that it has a short generation time (~15 generations/year or 0.07 years/generation, (Turelli &
543 Hoffmann, 1991)) but a relatively modest effective population size (i.e., historical N_e of
544 roughly 10^6 (Arguello et al., 2019)). Based on these figures, $8N_e$ would correspond to 0.56My,
545 much shorter than the equivalent time in humans, which makes it more likely that broader
546 environmental conditions remain constant enough to allow for the maintenance of balanced
547 polymorphisms in this species. In addition, extremely short generation times such as those in
548 *Drosophila* provide further scope for balancing selection due to seasonal fluctuations in the

environment (Wittmann et al., 2023), which is an unlikely avenue for balancing selection in species with long generation times.

The effective population size also has other direct and indirect effects on the presence and detectability of balanced polymorphisms, beyond the timescales discussed above. Large N_e leads to reduced drift and concomitantly more effective selection to maintain balanced polymorphisms and should therefore (all else being equal), increase the proportion of balanced loci that remains long-term polymorphic. On the other hand, N_e also has adverse effects on the signatures at linked neutral polymorphism. Species with large N_e tend to have larger effective recombination rates (larger $N_e r$), owing trivially to the effect of larger N_e but also due to the fact that large-N_e species often have relatively streamlined genomes and accordingly higher per-nucleotide recombination rates (r) (Lynch, 2006). Elevated recombination will erode the balancing selection signal of intermediate frequencies at neutral sites that are closely linked to selected variants and thereby make the signature sparser and inference harder. This effect can be compensated by a higher mutation rate (μ), which favours the generation of linked polymorphism around the balanced site. Thus, for a given N_e, the power to detect balancing selection is determined by the ratio of mutation to recombination rate, μ/r.

Demographic history and comparison to previous power estimates

It is worth noting that while we have only simulated simple scenarios of constant population size, more complex demographic histories are likely to substantially affect detection of balanced polymorphisms. Changes in the effective population size over time, such as sudden contractions (bottlenecks) and subsequent expansions, can have beneficial effects on detection. Power can be improved because a bottleneck in the not-too-distant past will result in increased rates of coalescence (a more recent average TMRCA) at neutral regions that are unlinked to sites under balancing selection, and thus on average tend to cause a leftward-shifted SFS at those regions. Peaks of intermediate frequencies and older TMRCAs at sites linked to balanced loci stand out more clearly from such genomic backgrounds of neutral diversity. It is likely that this phenomenon has facilitated the identification of balanced loci in humans, where migration out of Africa has resulted in a series of bottlenecks and founder events that have systematically shifted the neutral SFS.

The small N_e and specific demographic history of humans likely explains why previous analyses of the power of current inference methods, including the NCD statistic used here

581 (Bitarello et al., 2018) or BalLeRMix (Cheng & DeGiorgio, 2020), have provided higher
582 estimates of power (reaching over 80% under some conditions) than the ones we report here
583 for our human-like 'reference' parameters. While methods like NCD or BalLeRMix are general
584 in scope, they have typically been benchmarked using simulated data that follow a
585 demographic history and a mutation-to-recombination rate ratio that is characteristic of human
586 populations (e.g., 2.5 in Bitarello et al. (2018) and sampled between 2.5 and 0.83 in Cheng and
587 DeGiorgio (2020)), and relatively old polymorphisms (*e.g.* 5 million years in Bitarello et al.
588 (2018)and up to 650,000 generations with $N_e = 10^4$ in Cheng and DeGiorgio (2020), which
589 corresponds to 16.25 million years with the generation time of 25 years used in Bitarello et al.
590 (2018)). Given the specific assumptions made in the original method publications and the
591 substantial variation in power as a function of the parameters used, we reiterate that it is
592 imperative to combine any application of existing methods to a careful power analysis,
593 parameterised as much as possible to fit the study organism and population (Bitarello et al.,
594 2018, 2023).

595

596 **Methodological notes on NCD**

597 We have deliberately focused our analyses on a single approach, NCD (but see Supplementary
598 Material for additional analyses with BalLeRMix), to evaluate how different population genetic
599 parameters determine the efficacy of balancing selection and the signals it generates. It is not
600 our aim to compare different methods, all of which essentially exploit the same signals in the
601 data and will hence be affected similarly by the parameters we explore. Nevertheless, we will
602 briefly discuss a few specificities of our application of NCD that are expected to have small
603 effects on the absolute power we estimate.

604 First, in our analyses, we use the actual equilibrium frequency, which is known in our
605 simulations, as the NCD 'target frequency' (i.e., the frequency value from which NCD
606 measures the mean squared distance of allele frequencies in a window). Doing so will have a
607 small positive effect on power, compared to the case of an empirical application where the
608 target frequency is not known, and is therefore set to an arbitrary intermediate value. Previous
609 analyses (Bitarello et al., 2018), however, have shown that power is minimally impacted by the
610 exact frequency value of the target and additional analyses performed on our simulations
611 confirm this (Supplementary Material).

612 On the other hand, two other specificities of our application of NCD are likely to have
613 a small negative effect on power. First, we use a variant of NCD, NCD2, that takes into account
614 the SFS of derived alleles, including substitutions. Because we do not simulate an outgroup,
615 however, NCD2 is not as powerful here as the approach would be in a real-world application
616 where the SFS also includes fixed sites that result from substitutions inferred from outgroup
617 data (see Methods). Second, we use NCD with a fixed window size of 3Kb, independently of
618 the recombination rate. In cases where recombination rate is high and regions of elevated
619 polymorphism are narrow, calculating the SFS over a smaller interval can improve power,
620 although this is offset by the fact that the small numbers of SNPs in a narrow focal window
621 generate noise that negatively impacts the inference (Supplementary Material). Again, real-
622 world applications should be preceded by power analyses that explore these effects and can
623 help to optimise the analysis.

624 **General considerations on differences between species**

625 Our analyses make it clear that the prospects of identifying targets of balancing selection are
626 expected to vary significantly and systematically among species. Thus, species will harbour
627 variable numbers of long-term balanced loci, depending on the effective sizes of their
628 populations. Furthermore, the effective size, combined with mutation and recombination rates
629 will determine how difficult these sites are to identify as targets of balancing selection. In
630 humans, for example, the small effective population size means that only a limited number of
631 balanced polymorphisms are expected to have remained polymorphic over sufficiently long
632 periods of time. On the other hand, the same low effective population size, together with the
633 high ratio of mutation to recombination rate ($\mu/r = 1.2 \times 10^{-8}/1.3 \times 10^{-8} \sim 1$ (Pritchard,
634 2023 Chap. 1.1)), favour detection of such loci. In contrast, *D. melanogaster*, and possibly
635 other insects, presents a much harder case for identification, as the recombination rate is high
636 relative to the mutation rate ($\mu/r \sim 3 \times 10^{-9}/3 \times 10^{-8} = 0.1$ (Wang et al., 2023)), which
637 our analyses suggest should reduce power by roughly four-fold compared to a species with
638 human-like mutation and recombination parameters (mean power across simulated parameters
639 in Fig. 4A, 0.14 vs. 0.03).

640 Of course, as discussed above, there are other important factors at work that affect the
641 relative frequency of detectable balanced loci in the two species, such as the time required for
642 long-term balancing selection, the scope for seasonal selection, effects of N_e on the
643 maintenance of polymorphism on the one hand and its detectability on the other, and species-

644 specific demographic histories. Overall, there is little doubt that the detection of balancing
645 selection in *Drosophila* is challenging, which suggests that there could be many targets of
646 balancing selection in *D. melanogaster* that are invisible to genomic scans for balancing
647 selection, and this may also be the case for many other species with similar population
648 parameters. Indeed, a genome-wide screen for balancing selection in fruit flies combining
649 Theta and Tajima's *D* identified roughly 100 candidate balanced loci (Croze et al., 2017),
650 compared to over 1000 candidates in humans (Bitarello et al., 2018). Although more
651 sophisticated methods may improve power in *Drosophila*, the notion that large numbers of
652 balanced polymorphisms might exist and be missed in genomic scans is in line with the high
653 levels of standing genetic variation for major fitness components in this species, which exceeds
654 the variation explainable under mutation-selection-drift balance (Charlesworth, 2015; Sharp &
655 Agrawal, 2018). The excess variance in life-history traits might be attributable to balanced
656 polymorphisms segregating in *D. melanogaster* populations. While it seems plausible that at
657 least some of these balanced polymorphisms would be detected if detection methods were more
658 powerful, others could remain invisible if they are relatively young (relative to $4N_e$). It is
659 difficult to assess these alternatives in the absence of more powerful approaches, and similar
660 conclusions may apply to many other species with equally unfavourable population parameters.

661

662 # Methods

663 **Population genetic simulations of balancing selection**

664 We used simulations to assess which regions of the parameter space generate genomic
665 signatures of balancing selection that can be detected using current approaches. Simulations
666 were run in SLiM (version 4.0.1, (Haller et al., 2019)), a forward simulation framework for
667 population genetics. In line with the analytical derivations above, we implemented
668 overdominance (OD), antagonistic pleiotropy (AP) and sexually antagonistic selection (SA),
669 in addition to simulations under neutrality. In the simulations with selection, we simulated a
670 single focal site, embedded centrally in a chromosome of 10,000bp, large enough for neutral
671 sites in the periphery to be unaffected by selection at the central site. To implement the three
672 selection mechanisms, we used custom functions (see SLiM scripts in Supplementary Material)
673 to run a non-Wright-Fisher model for populations of constant size with non-overlapping
674 generations. In each generation, the code generated offspring by randomly drawing parents
675 based on their fitness values, calculated according to their genotype at the locus under selection

676 and the selection mechanism and parameters (selection coefficients, dominance, cf. Tab. S1 in
677 the Appendix) being simulated. We used the same code for reproduction in the neutral
678 simulations, but set the fitness of all individuals to 1. In all simulations the default sex ratio of
679 1:1 was used.

680 At the start of each simulation, a mutation at the focal site was introduced on a single
681 chromosome, randomly chosen within the population. The mutation would be of type A_1 or A_2
682 (see Tab. A1, *e.g.*, either male- or female-beneficial for SA) with equal probability. To reduce
683 the number of simulation runs where the polymorphism was quickly lost due to drift when rare,
684 we initially subjected the mutant allele to directional positive selection with $s = 0.5$ and $h = 1/2$
685 until the mutant allele frequency reached 5% (or the mutant allele was lost). This selective
686 boost facilitates the establishment of the polymorphism, but due to its old age and short
687 duration should have negligible influence on the patterns of linked variation. For
688 polymorphisms that reached a frequency of 5%, simulations then switched from directional to
689 balancing selection with a specified mechanism and parameter combination. We recorded the
690 frequency of the mutant (derived) allele and the genealogy of all chromosomes in the
691 population at 16000, 32000, 160000 and 320000 generations.

692 To optimise the runtime of simulations, we did not include neutral mutations in any
693 of the SLiM simulations. Instead, we recorded the full chromosome-wide genealogies from
694 each simulation (using 'tree-sequence recording', (Haller et al., 2019)). We then used
695 'recapitation' in pyslim (version 1.0.4 (Ralph et al., 2023)) to generate the preceding
696 genealogical relationships of the starting chromosomes up to their most recent common
697 ancestor. Finally, we added neutral mutations on the complete recapitated genealogies with
698 msprime (version 1.2.0, (Baumdicker et al., 2022)).

699

Table 3. Parameter values used in the simulations. We highlight in bold the values used in our "reference" set of simulations, which approximate the parameter values estimated in the literature for humans.

Selection model	Overdominance (OD) Antagonistic pleiotropy (AP) Sexual antagonism (SA)
Selection coefficients (s)	s_{weak} = {0.01, 0.02,0.05,0.1} s_{strong} = {0.1,0.2,0.5,1}

Dominance coefficients (*h*, only defined for AP and SA)	0.25 0.5
Time of onset of selection (*T*, in generations)	**8Ne** 16Ne
(Effective) population size (N=N_e)	2,000 **20,000**
Recombination rate (r)	10^{-7} **10^{-8}**
Mutation rate (μ)	10^{-7} **10^{-8}**

The parameter values used in our simulations are summarised in Table 3. We explored selection and dominance coefficients to match the analytical results. For the species-specific parameters of effective population size (N_e), recombination rate (r) and mutation rate (μ) were chosen to cover a broad range of relevant scenarios. Thus, we simulated population sizes of $N = N_e = 2{,}000$ and $N = N_e = 20{,}000$ and neutral mutation at rates of $\mu = 10^{-7}$ and $\mu = 10^{-8}$ per base per generation, such that the expected genetic diversity ($\pi = 4N_e\mu$) ranges from $8\text{x}10^{-5}$ to $8\text{x}10^{-3}$. The highest expected genetic diversity (for $N_e = 20{,}000$ and $\mu = 10^{-7}$) approaches that of *Drosophila melanogaster* ($\pi = 10^{-2}$ (Leffler et al., 2012)), while the combination of $N_e = 20{,}000$ and $\mu = 10^{-8}$ generates diversity similar to that estimated in humans ($\pi = 8\text{x}10^{-4}$ (Pritchard, 2023 Chap.1.1)). We simulated recombination rates of $r=10^{-7}$ and $r = 10^{-8}$ per base per generation to include combinations of $\mu = r = 10^{-8}$, which are similar to the values estimated in humans (Pritchard, 2023 Chap. 1.1), as well as a ratio of mutation to recombination rate similar to that estimated in *Drosophila melanogaster* ($\mu/r = 10^{-8}/10^{-7} = 0.1$, (Wang et al., 2023)) and a scenario where mutation rate exceeds recombination rate ($\mu/r = 10^{-7}/10^{-8} = 10$), which should facilitate the detection of signatures of selection. Finally, we used two values for the onset of selection (or age of the polymorphism), $T = 8N_e$ and $T = 16N_e$ (16,000 and 32,000 generations for $N_e = 2{,}000$ and 160,000 and 320,000 generations for $N_e = 20{,}000$). These values roughly correspond to the definitions of long-term and ultra long-term balancing selection (Bitarello et al. 2023). Both significantly exceed the expected coalescence time (time to the

720 most recent common ancestor, TMRCA) under neutrality, $4N_e$. The probability of observing a
721 TMRCA larger than $8N_e$ generations under neutrality is approximately 0.054, while that of a
722 TMRCA larger than $16N_e$ generations is approximately 0.001 (Wakeley 2009, p.77 eqn 3.27).
723 To explore the effect of the strength of selection, equilibrium frequency and model of
724 selection on the maintenance of the selected polymorphism (Figs. S1-S4) and the inference of
725 balancing selection (Fig. 3), we ran simulations for a reference set with human-like parameters
726 (bold values in Tab. 1, N_e=20,000, $\mu = r = 10^{-8}$, $T = 8N_e$ =160,000 generations) under OD, AP
727 and SA across combinations of selection coefficients s_i and s_j drawn from two 4-by-4 grids
728 with weaker selection {0.01,0.02,0.05,0.1} and stronger selection {0.1,0.2,0.5,1.0}. To explore
729 the effects of mutation rate, recombination rate, selection onset time and N_e, (Fig. 5), we ran
730 similar sets of simulations for OD only, but changing the values of each of these parameters
731 relative to the reference set of simulations (i.e. $\mu = 10^{-7}$, $r = 10^{-7}$, $T = 16N_e$ = 320,000
732 generations, N_e = 2,000).
733 In simulations to assess the maintenance of polymorphisms, we ran 100 replicate
734 simulations in which the mutation successfully established (*i.e.*, reached 5%) for each
735 parameter combination. To assess the power to detect signatures of balancing selection, we
736 complemented these with additional simulations to obtain sets of 100 replicate simulations for
737 each parameter combination in which the polymorphism was maintained.
738
739 **Power of inference of balancing selection**

740 We used the NCD (non-central deviation) method (Bitarello et al., 2018) to infer balancing
741 selection on the simulated data (but for complementary results using BalLeRMix (Cheng &
742 DeGiorgio, 2020), see Supplementary Material). Like most other current inference methods,
743 NCD detects regions under balancing selection due to their signal of excess intermediate-
744 frequency polymorphism. NCD for a genomic window with n informative sites is defined as:

745 $$NCD(tf) = \sqrt{\frac{\sum_{i=1}^{n}(p_i - tf)^2}{n}}$$

746 where p_i is the allele frequency at the ith informative site, and tf is the target frequency (set
747 as an intermediate value *e.g.*, 0.3, 0.4 or 0.5). The presence of a balanced site in the focal region
748 with allele frequencies close to the target frequency then results in a low value of the NCD
749 statistic, compared to a region under neutrality or purifying selection.

750 In our application of this approach, we used a refined version of the statistic, NCD2.
751 This statistic gains power by also considering fixed differences (relative to an outgroup) in the
752 site frequency spectrum (Bitarello et al., 2018), thus exploiting the signal of increased
753 polymorphism and lower rate of substitution in regions under balancing selection. Although
754 we have not simulated an outgroup, we can differentiate ancestral and derived alleles in our
755 simulated genotypes and then manually add an 'outgroup' individual to the VCF file from our
756 simulations, for which all sites are homozygous ancestral. Therefore, any sites that are fixed
757 for the derived allele among the simulated (ingroup) individuals will be counted as fixed
758 differences in the NCD2 calculation. We note that this approach adds less additional power
759 than a real outgroup would, as the latter would also contribute fixed differences that arise due
760 to substitutions in the outgroup lineage.
761 To assess the statistical significance of the NCD2 values, we generated a null
762 distribution by running 1000 neutral simulations for each combination of relevant parameters
763 (N_e, μ and r) and calculating NCD2 for each. We then calculated a P-value for an NCD2 value
764 from a simulation with selection by determining the proportion of NCD2 values from the null
765 distribution (with corresponding N_e, μ and r) that was lower than the NCD2 value with
766 selection. Power was then calculated as the proportion of the 100 simulations generated under
767 a particular set of parameters for which $P \leq 0.01$.
768
769 **Statistical analysis of power estimates**
770 We fitted binomial Generalised Linear Models (GLMs) to our power data to directly model the
771 probability of detecting or failing to detect balancing selection as a function of the parameters
772 that were varied in our simulations. The models used a logit link function and were fitted to the
773 number of successful and unsuccessful instances of detection as the response variable. The
774 significance of model terms was determined with an analysis of deviance, a series of sequential
775 likelihood ratio tests based on the Chi^2 distribution. This analysis further makes it possible to
776 assess the importance of each model term through its deviance (the amount of variability in the
777 response explained by a model term) relative to the null deviance, the total amount of variability
778 in the response values analysed. We note that all models showed a degree of overdispersion,
779 as indicated by the residual deviance exceeding the number of residual degrees of freedom
780 several-fold. However, running alternative models that accommodate overdispersion (as a
781 'quasibinomial' model, combined with F-tests) produced qualitatively identical results.

These GLMs fit a linear (additive) model of detection probability (π) as a function of the simulation parameters. The model is fitted on the log-odds scale (where odds are $\pi/(1-\pi)$) that accommodates the fact that probabilities are bounded by 0 and 1. The estimated model parameters, back-transformed to the odds-scale, provide information on the proportional change in the detection probability that is independent of the value of π (while the change in the actual probability due to a change in a predictor variable changes over the range of π).

Acknowledgements

We would like to thank Filip Ruzicka and members of the Alliance EvolGen lab group for helpful discussions. MR, AA and DB were supported by grants from the UK Biotechnology and Biological Sciences Research Council, BBSRC (award BB/W007703/1 to MR and AA) and the Leverhulme Trust (award RPG-2021-414 to MR and AA). The work was initiated and developed during mutual visits between London and Melbourne by MR, AA and TC, made possible by an Australia Partnering award from the BBSRC (award BB/T019921/1 to MR and TC).

Bibliography

Allison, A. C. (1954). The distribution of the sickle-cell trait in East Africa and elsewhere, and its apparent relationship to the incidence of subtertian malaria. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *48*(4), 312–318. https://doi.org/10.1016/0035-9203(54)90101-7

Arguello, J. R., Laurent, S., & Clark, A. G. (2019). Demographic History of the Human Commensal Drosophila melanogaster. *Genome Biology and Evolution*, *11*(3), 844–854. https://doi.org/10.1093/gbe/evz022

Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P., Jacq, C., Jensen, A. J., Johnston, S. E., Karlsson, S., Kent, M., Moen, T., Niemelä, E., Nome, T., Næsje, T. F., Orell, P., Romakkaniemi, A., Sægrov, H., Urdal, K., … Primmer, C. R. (2015).

Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature*, *528*(7582), 405–408. https://doi.org/10.1038/nature16062

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., … Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3), iyab229. https://doi.org/10.1093/genetics/iyab229

Bitarello, B. D., Brandt, D. Y. C., Meyer, D., & Andrés, A. M. (2023). Inferring Balancing Selection From Genome-Scale Data. *Genome Biology and Evolution*, *15*(3), evad032. https://doi.org/10.1093/gbe/evad032

Bitarello, B. D., De Filippo, C., Teixeira, J. C., Schmidt, J. M., Kleinert, P., Meyer, D., & Andrés, A. M. (2018). Signatures of Long-Term Balancing Selection in Human Genomes. *Genome Biology and Evolution*, *10*(3), 939–955. https://doi.org/10.1093/gbe/evy054

Chao, L., & Carr, D. E. (1993). The molecular clock and the relationship between population size and generation time. *Evolution*, *47*(2), 688–690. https://doi.org/10.1111/j.1558-5646.1993.tb02124.x

Charlesworth, B. (2015). Causes of natural variation in fitness: Evidence from studies of Drosophila populations. *Proceedings of the National Academy of Sciences*, *112*(6), 1662–1669. https://doi.org/10.1073/pnas.1423275112

Charlesworth, B., & Charlesworth, D. (2010). *Elements of Evolutionary Genetics*. Roberts and Company Publishers.

Charlesworth, B., & Hughes, K. A. (2000). The Maintenance of Genetic Variation in Life-History Traits. In *Evolutionary Genetics: From molecules to morphology*.

Cheng, X., & DeGiorgio, M. (2020). Flexible Mixture Model Approaches That Accommodate Footprint Size Variability for Robust Detection of Balancing Selection. *Molecular Biology and Evolution*, *37*(11), 3267–3291. https://doi.org/10.1093/molbev/msaa134

Connallon, T., & Chenoweth, S. F. (2019). Dominance reversals and the maintenance of genetic variation for fitness. *PLOS Biology*, *17*(1), e3000118. https://doi.org/10.1371/journal.pbio.3000118

Connallon, T., & Clark, A. G. (2014). Balancing Selection in Species with Separate Sexes: Insights from Fisher's Geometric Model. *Genetics*, *197*(3), 991–1006. https://doi.org/10.1534/genetics.114.165605

Connallon, T., & Czuppon, P. (2026). Additive genetic variance for fitness maintained by balancing selection. *Evolution, in Press*.

Crow, J. F. (1952). Dominance and Overdominance. In *Heterosis: A Record of Researches Directed Toward Explaining and Utilizing the Vigor of Hybrids*. Iowa State College Press.

Crow, J. F. (1987). Muller, Dobzhansky, and overdominance. *Journal of the History of Biology*, *20*(3), 351–380. https://doi.org/10.1007/BF00139460

Croze, M., Wollstein, A., Božičević, V., Živković, D., Stephan, W., & Hutter, S. (2017). A genome-wide scan for genes under balancing selection in Drosophila melanogaster. *BMC Evolutionary Biology*, *17*(1), 15. https://doi.org/10.1186/s12862-016-0857-z

Curtsinger, J. W., Service, P. M., & Prout, T. (1994). Antagonistic Pleiotropy, Reversal of Dominance, and Genetic Polymorphism. *The American Naturalist*. https://doi.org/10.1086/285671

DeGiorgio, M., Lohmueller, K. E., & Nielsen, R. (2014). A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data. *PLOS Genetics*, *10*(8), e1004561. https://doi.org/10.1371/journal.pgen.1004561

Dobzhansky, T. (1955). A Review of Some Fundamental Concepts Problems of Population Genetics. *Cold Spring Harbor Symposia on Quantitative Biology*, *20*, 1–15.

Fijarczyk, A., & Babik, W. (2015). Detecting balancing selection in genomes: Limits and prospects. *Molecular Ecology*, *24*(14), 3529–3545. https://doi.org/10.1111/mec.13226

Fisher, R. A. (1922). On the Dominance Ratio. *Proceedings of the Royal Society of Edinburgh*, *42*, 321–341. https://doi.org/10.1017/S0370164600023993

Fitzpatrick, M. J., Feder, E., Rowe, L., & Sokolowski, M. B. (2007). Maintaining a behaviour polymorphism by frequency-dependent selection on a single gene. *Nature*, *447*(7141), 210–212. https://doi.org/10.1038/nature05764

Glémin, S. (2021). Balancing selection in self-fertilizing populations. *Evolution; International Journal of Organic Evolution*, *75*(5), 1011–1029. https://doi.org/10.1111/evo.14194

Grieshop, K., Ho, E. K. H., & Kasimatis, K. R. (2024). Dominance reversals: The resolution of genetic conflict and maintenance of genetic variation. *Proceedings of the Royal Society B: Biological Sciences*, *291*(2018), 20232816. https://doi.org/10.1098/rspb.2023.2816

Haldane, J. B. S., & Jayakar, S. D. (1963). Polymorphism due to selection depending on the composition of a population. *Journal of Genetics*, *58*(3), 318–323. https://doi.org/10.1007/BF02986301

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., & Ralph, P. L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, *19*(2), 552–566. https://doi.org/10.1111/1755-0998.12968

Hedrick, P. W. (1999). Antagonistic pleiotropy and genetic polymorphism: A perspective. *Heredity*, *82*(2), 126–133. https://doi.org/10.1038/sj.hdy.6884400

Hedrick, P. W. (2012). What is the evidence for heterozygote advantage selection? *Trends in Ecology & Evolution*, *27*(12), 698–704. https://doi.org/10.1016/j.tree.2012.08.012

Isildak, U., Stella, A., & Fumagalli, M. (2021). Distinguishing between recent balancing selection and incomplete sweep using deep neural networks. *Molecular Ecology Resources*, *21*(8), 2706–2718. https://doi.org/10.1111/1755-0998.13379

Johnston, S. E., Gratten, J., Berenos, C., Pilkington, J. G., Clutton-Brock, T. H., Pemberton, J. M., & Slate, J. (2013). Life history trade-offs at a single locus maintain sexually selected genetic variation. *Nature*, *502*(7469), 93–95. https://doi.org/10.1038/nature12489

Karageorgi, M., Lyulina, A. S., Bitter, M. C., Lappo, E., Greenblum, S. I., Mouza, Z. K., Tran, C. T., Huynh, A. V., Oken, H., Schmidt, P., & Petrov, D. A. (2025). Beneficial reversal of dominance maintains a large-effect resistance polymorphism under fluctuating insecticide selection. *Nature Ecology & Evolution*, *9*(11), 2145–2160. https://doi.org/10.1038/s41559-025-02853-x

Kidwell, J. F., Clegg, M. T., Stewart, F. M., & Prout, T. (1977). Regions of stable equilibria for models of differential selection in the two sexes under random mating. *Genetics*, *85*(1), 171–183. https://doi.org/10.1093/genetics/85.1.171

Kimura, M. (1968). Evolutionary Rate at the Molecular Level. *Nature*, *217*(5129), 624–626. https://doi.org/10.1038/217624a0

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511623486

King, J. L., & Jukes, T. H. (1969). Non-Darwinian Evolution. *Science*, *164*(3881), 788–798. https://doi.org/10.1126/science.164.3881.788

Klein, J., Sato, A., Nagl, S., & O'hUigín, C. (1998). Molecular trans-species polymorphism. *Annual Review of Ecology, Evolution, and Systematics*, *29*(Volume 29, 1998), 1–21. https://doi.org/10.1146/annurev.ecolsys.29.1.1

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. *Nature*, *304*(5925), 412–417. https://doi.org/10.1038/304412a0

Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., Andolfatto, P., & Przeworski, M. (2012). Revisiting an old riddle: What determines genetic diversity levels within species? *PLoS Biology*, *10*(9), e1001388. https://doi.org/10.1371/journal.pbio.1001388

Levene, H. (1953). Genetic Equilibrium When More Than One Ecological Niche is Available. *The American Naturalist*. https://doi.org/10.1086/281792

Lewontin, R. C., & Hubby, J. L. (1966). 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*, *54*(2), 595–609. https://doi.org/10.1093/genetics/54.2.595

Lewontin, R. C. (with Internet Archive). (1974). *The genetic basis of evolutionary change*. New York, Columbia University Press. http://archive.org/details/geneticbasisofev00lewo

Lynch, M. (2006). The Origins of Eukaryotic Gene Structure. *Molecular Biology and Evolution*, *23*(2), 450–468. https://doi.org/10.1093/molbev/msj050

Nei, M., & Roychoudhury, A. K. (1973). Probability of Fixation and Mean Fixation Time of an Overdominant Mutation. *Genetics*, *74*(2), 371–380. https://doi.org/10.1093/genetics/74.2.371

Pritchard, J. K. (2023). *An Owner's Guide to the Human Genome* (Release 1.0).

Prout, T. (2000). How well does opposing selection maintain variation? In *Evolutionary Genetics: From molecules to morphology* (pp. 157–181). Cambridge University Press.

Ralph, P., Jeffery, B., Wong, Y., Ashander, J., Patterson, G., Kelleher, J., R, M., Kern, A., Tittes, S., Huang, X., & Galloway, J. (2023). *tskit-dev/pyslim: 1.0.4* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.8205346

Reid, J. (2022). Intrinsic emergence and modulation of sex-specific dominance reversals in threshold traits. *Evolution*, *76*, 1924–1941.

Rice, S. H. (2004). *Evolutionary Theory: Mathematical and Conceptual Foundations*. Sinauer.

Robertson, A. (1962). Selection for Heterozygotes in Small Populations. *Genetics*, *47*(9), 1291–1300. https://doi.org/10.1093/genetics/47.9.1291

Rose, M. R. (1982). Antagonistic pleiotropy, dominance, and genetic variation. *Heredity*, *48*(1), 63–78. https://doi.org/10.1038/hdy.1982.7

Rusuwa, B. B., Chung, H., Allen, S. L., Frentiu, F. D., & Chenoweth, S. F. (2022). Natural variation at a single gene generates sexual antagonism across fitness components in Drosophila. *Current Biology*, *32*(14), 3161–3169. https://doi.org/10.1016/j.cub.2022.05.038

Ruzicka, F., Zwoinska, M. K., Goedert, D., Kokko, H., Li Richter, X.-Y., Moodie, I. R., Nilén, S., Olito, C., Svensson, E. I., Czuppon, P., & Connallon, T. (2025). A century of theories of balancing selection. *Biological Reviews*. https://doi.org/10.1111/brv.70103

Sharp, N. P., & Agrawal, A. F. (2018). An experimental test of the mutation-selection balance model for the maintenance of genetic variance in fitness components. *Proceedings. Biological Sciences*, *285*(1890), 20181864. https://doi.org/10.1098/rspb.2018.1864

Siewert, K. M., & Voight, B. F. (2017). Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. *Molecular Biology and Evolution*, *34*(11), 2996–3005. https://doi.org/10.1093/molbev/msx209

Soni, V., Vos, M., & Eyre-Walker, A. (2022). A new test suggests hundreds of amino acid polymorphisms in humans are subject to balancing selection. *PLOS Biology*, *20*(6), e3001645. https://doi.org/10.1371/journal.pbio.3001645

Turelli, M., & Hoffmann, A. A. (1991). Rapid spread of an inherited incompatibility factor in California Drosophila. *Nature*, *353*(6343), 440–442. https://doi.org/10.1038/353440a0

Wang, Y., McNeil, P., Abdulazeez, R., Pascual, M., Johnston, S. E., Keightley, P. D., & Obbard, D. J. (2023). Variation in mutation, recombination, and transposition rates in Drosophila melanogaster and Drosophila simulans. *Genome Research*, *33*(4), 587–598. https://doi.org/10.1101/gr.277383.122

Williamson, S., Fledel-Alon, A., & Bustamante, C. D. (2004). Population Genetics of Polymorphism and Divergence for Diploid Selection Models With Arbitrary Dominance. *Genetics*, *168*(1), 463–475. https://doi.org/10.1534/genetics.103.024745

Wittmann, M. J., Mousset, S., & Hermisson, J. (2023). Modeling the genetic footprint of fluctuating balancing selection: From the local to the genomic scale. *Genetics*, *223*(4), iyad022. https://doi.org/10.1093/genetics/iyad022

Wright, S. (1942). Statistical genetics and evolution. *Bulletin of the American Mathematical Society*, *48*(4), 223–246. https://doi.org/10.1090/S0002-9904-1942-07641-5

Wright, S. (1945). The Differential Equation of the Distribution of Gene Frequencies. *Proceedings of the National Academy of Sciences*, *31*(12), 382–389. https://doi.org/10.1073/pnas.31.12.382

Zajitschek, F., & Connallon, T. (2018). Antagonistic pleiotropy in species with separate sexes, and the maintenance of genetic variation in life-history traits and fitness. *Evolution*, *72*(6), 1306–1316. https://doi.org/10.1111/evo.13493

Appendices

Appendix 1 – Conditions for internal frequency equilibria

We model balancing selection making three simplifying assumptions that should apply broadly to many (perhaps most) cases where this type of selection is present. First, we focus on populations of diploid, randomly mating (outbred) individuals (see Glémin (2021) for a comprehensive treatment of balancing selection in self-fertilizing species). Second, we consider loci with two major allele types, which we label A_1 (at a population frequency of p) and A_2 (frequency $1 - p$). Third, in scenarios involving antagonistic selection (i.e., antagonistic pleiotropy and sexual antagonism), we only consider scenarios of dominance that are amenable to balanced polymorphism, i.e. cases where each fitness component or sex either lacks dominance (fitness effects are additive in each context of selection) or there is a beneficial reversal of dominance, in which the costs of expressing the "wrong" allele for a given context of selection are partially or completely recessive (reviewed in Connallon and Chenoweth 2019; Grieshop et al. 2024). Our exposition draws upon previous studies of balancing selection theory (*e.g.*, Robertson 1962; Kidwell et al. 1977; Curtsinger et al. 1994; Connallon and Clark 2013), with Table 1 providing an outline of the fitness values associated with each genotype and for each model of selection.

Table S1. Fitness per genotype for each model of selection

	Genotype		
	A_1A_1	A_1A_2	A_2A_2
Overdominance	$1 - s_1$	1	$1 - s_2$
Antagonistic pleiotropy			
Fitness component 1	1	$1 - t_1h$	$1 - t_1$
Fitness component 2	$1 - t_2$	$1 - t_2h$	1
Overall Fitness	$1 - t_2$	$(1 - t_1h)(1 - t_2h)$	$1 - t_1$
Sexual antagonism			

Female fitness	1	$1 - s_f h_f$	$1 - s_f$
Male fitness	$1 - s_m$	$1 - s_m h_m$	1

Parameters for the three models fall within the following ranges: overdominance: $1 \geq s_1, s_2 > 0$; antagonistic pleiotropy: $1 \geq t_1, t_2 > 0$; sexual antagonism: $1 \geq s_f, s_m > 0$. Dominance coefficients are within the range $0.5 \geq h, h_f, h_m \geq 0$, which includes additive effects (h = 0.5) and beneficial reversals of dominance (0.5 > h). Note that many of our analyses, we focus on scenarios where fitness effects are small, *i.e.*.: $1 \gg s_1, s_2 > 0$; $1 \gg t_1, t_2 > 0$; and $1 \gg s_f, s_m > 0$. Note that in the Appendices, we use slightly different notation to distinguish selection parameters of the OD, AP, and SA models (s_1/s_2, t_1/s_2, and s_f/s_m, respectively). For simplicity, and with no loss of generality, we use a single notation, s_1/s_2, in the main text.

Overdominant selection (OD)

With OD, the heterozygotes for the locus (A_1A_2) have the highest fitness and A_1A_1 and A_2A_2 homozygotes suffer fitness reductions of s_1 and s_2, respectively (Table 1). Evolutionary dynamics are captured by the following difference equation describing allele frequency change across a single generation:

$$\Delta p = \frac{(s_1 + s_2)p(1-p)(\hat{p} - p)}{\overline{w}} \approx (s_1 + s_2)p(1-p)(\hat{p} - p) \qquad (1)$$

where $\underline{w} = p^2(1 - s_1) + 2p(1-p) + (1-p)^2(1 - s_2)$ denotes mean fitness and:

$$\hat{p} = \frac{s_2}{s_1 + s_2}$$

is the equilibrium frequency for the A_1 allele. The final approximation in eq. (1) applies under weak selection ($0 < s_1, s_2 << 1$). Note that the entire range of parameter values in this model generates balancing selection (*i.e.*, $0 < \hat{p} < 1$ for the plausible parameter range, $0 < s_1, s_2 < 1$).

Antagonistic pleiotropy (AP)

Following previous models (e.g. Curtsinger et al., 1994; Rose, 1982), we consider a locus with antagonistic pleiotropic effects on two major fitness components. Selection is directional within each component (Table 1) and fitness components combine multiplicatively to determine overall fitness. The evolutionary dynamics are described by:

$$\Delta p = \frac{\left((t_1 + t_2)(1 - 2h) + 2t_1 t_2 h^2\right) p(1-p)(\hat{p} - p)}{\overline{w}}$$

$$\approx \left((t_1 + t_2)(1 - 2h) + 2t_1 t_2 h^2\right) p(1-p)(\hat{p} - p) \qquad (3)$$

where $\overline{w} = p^2(1 - t_2) + 2p(1-p)(1 - t_1 h)(1 - t_2 h) + (1-p)^2(1 - t_1)$ and the polymorphic equilibrium (whenever it is valid) is:

$$\hat{p} = \frac{t_1(1-h) - t_2 h + t_1 t_2 h^2}{(t_1 + t_2)(1 - 2h) + 2t_1 t_2 h^2} \qquad (4)$$

The equilibrium is valid under the following condition:

$$\frac{t_2 h}{1 - h + t_2 h^2} < t_1 < \frac{t_2(1-h)}{h(1 - t_2 h)} \qquad (5)$$

and otherwise, selection on total fitness is directional and favours fixation of either A_1 or A_2.

Sexually antagonistic selection (SA)

Following Kidwell et al. (1977), sexually antagonistic alleles are under opposing directional selection in males and females, characterised by a selection and dominance coefficient for each sex (Table 1). Letting $p = (p_f + p_m)/2$ represent the overall frequency of A_1 in the population, where p_f and p_m represent the respective frequencies in female and male gametes contributing to fertilization, the exact allele frequency dynamics are given by $\Delta p = \frac{1}{2}\left(\Delta p_f + \Delta p_m\right)$, where:

$$\Delta p_f = \frac{p_m p_f + \left(p - p_m p_f\right)\left(1 - s_f h_f\right)}{p_m p_f + 2\left(p - p_m p_f\right)\left(1 - s_f h_f\right) + \left(1 - p_f\right)(1 - p_m)\left(1 - s_f\right)} - p_f \qquad (6a)$$

and

$$\Delta p_m = \frac{p_m p_f (1 - s_m) + \left(p - p_m p_f\right)(1 - s_m h_m)}{p_m p_f (1 - s_m) + 2\left(p - p_m p_f\right)(1 - s_m h_m) + \left(1 - p_f\right)(1 - p_m)} - p_m \qquad (6b)$$

Within the range $0 \leq h = h_f = h_m \leq 0.5$ which we consider here, and assuming modest to weak selection ($0 < s_f, s_m << 1$), the allele frequency dynamics are closely approximated by:

$$\Delta p \approx \left(s(1 - 2h) + 2h^2 s_f s_m\right) p(1-p)(\hat{p} - p) \qquad (7)$$

(Connallon and Clark 2013), where $s = \frac{1}{2}\left(s_f + s_m\right)$, and $\hat{p}$ represents the polymorphic equilibrium when valid. With no dominance in either sex ($h = 0.5$), the equilibrium frequency (when it exists) is $\hat{p} = \frac{(s_f - s_m + s_f s_m)}{2 s_f s_m}$, and for cases involving dominance reversals ($h < 0.5$), the equilibrium is $\hat{p} \approx \frac{(s_f(1-h) - s_m h)}{(s_f + s_m)(1 - 2h)}$. The condition favouring a polymorphic equilibrium (based on the exact eqs. (6a-6b)) is:

$$\frac{s_m h_m}{1 - h_f + s_m h_m} < s_f < \frac{s_m(1 - h_m)}{h_f(1 - s_m)} \tag{8}$$

(Kidwell et al. 1977; Connallon and Chenoweth 2009).

If we assume that each selection coefficient is uniformly distributed with a mean of $\bar{s} = \frac{1}{2}s_{\max}$ (where $s_{\max}$ represents the maximum selection coefficient, $0 < s_{\max} \leq 1$), then the proportion of parameter space that results in balancing selection under AP will be:

$$f_{bal} = 1 - \frac{2h^2 s_{\max} + 2(1-h)\log\left(\frac{1-h}{1-h+h^2 s_{\max}}\right)}{s_{\max}^2 h^3} = \frac{1-2h}{1-h} + \frac{4h^3\bar{s}}{3(1-h)^2} + O(s_{\max}^2)$$

while that that results in balancing selection under SA is:

$$f_{bal} = 1 - \frac{2h s_{\max} + 2(1-h)\log\left(\frac{1-h}{1-h+h s_{\max}}\right)}{s_{\max}^2 h} = \frac{1-2h}{1-h} + \frac{4h^2\bar{s}}{3(1-h)^2} + O(s_{\max}^2)$$

Appendix 2 - Criteria for maintaining a long-term balanced polymorphism

A long-term polymorphism requires balancing selection to be both consistent over time and strong enough to offset the tendency of genetic drift to eliminate the polymorphism. Provided balancing selection is temporally consistent, then its capacity to maintain polymorphism can be quantified using standard diffusion theory (Rice, 2004; Wright, 1945). In the models presented above, the expected change in allele frequencies at a single locus under balancing selection (with change measured across a single generation) is:

$$M = \alpha p(1-p)(\hat{p} - p) \tag{9}$$

where p again refers to the frequency of A_1, p^ is the equilibrium under balancing selection, and α represents a model-specific constant that reflects the strength of balancing selection (these can be extracted from eqs. (1), (3) and (7), which yield: $\alpha = s_1 + s_2$ for the overdominance model, $\alpha = (t_1 + t_2)(1 - 2h) + 2t_1 t_2 h^2$ under antagonistic pleiotropy, and $\alpha = s(1 - 2h) + 2s_f s_m h^2$ under sexual antagonism). The variance in allele frequency change owing to drift is $V = \frac{p(1-p)}{2N_e}$, where N_e is the effective size of the population. Ignoring effects of recurrent mutation, the stationary distribution loci subject to balancing selection (*i.e.*, the long-term probability density function for allele frequency p) is:

$$f(p) = \frac{C}{p(1-p)} \exp\left((-2N_e\alpha(\hat{p} - p)^2)\right) \tag{10}$$

(Wright 1945; Robertson 1962; Rice 2004) where the constant C ensures that the probability density function integrates to one (*i.e.*, $C^{-1} = \int_0^1 [p(1-p)]^{-1} \exp(-2N_e\alpha(\hat{p} - p)^2)\, dp$).

Balancing selection effectively maintains polymorphism when the probability density near the equilibrium (p near $\hat{p}$) is high relative to the probability density near the boundaries of the distribution ($p = 1/2N_e$ and $p = 1 - 1/2N_e$). A measure of these relative densities (following Connallon & Clark, 2014) is given by:

$$R \approx \frac{1}{2N_e\hat{p}(1-\hat{p})\big(exp(-2N_e\alpha\hat{p}^2) + exp(-2N_e\alpha(1-\hat{p})^2)\big)} \tag{11}$$

for which balancing selection is effective when R is large ($R >> 1$), and drift dominates when R is small ($R < 1$). Using a cut-off of $R > 10$ to represent effectively strong balancing selection, eq. (11) implies the following minimum values of α ($a_{\min}$) are required to ensure the stability of a balanced polymorphism:

$$a_{\min} = \begin{cases} \dfrac{\log(20N_e\hat{p})}{2N_e\hat{p}^2} & \text{for } \hat{p} \leq 0.5 \\ \dfrac{\log\big(20N_e(1-\hat{p})\big)}{2N_e(1-\hat{p})^2} & \text{for } \hat{p} \geq 0.5 \end{cases} \tag{12}$$

From eq. (12), balancing selection should be effective at maintaining polymorphism when $\alpha > a_{\min}$, whereas drift will impede maintenance when α falls substantially below this threshold.

Figure 2 illustrates conditions for effective balancing selection across the spectrum of polymorphic equilibrium frequency states ($0 < \hat{p} < 1$) and for three different effective population sizes (N_e = 1,000; N_e = 10,000; N_e = 100,000). Two special cases of moderately strong balancing selection in a small population (N_e = 1,000 and α = 0.05) illustrate how the probability density of the stationary distribution is concentrated near $p = \hat{p}$ when $\alpha > a_{\min}$, whereas it accumulates near $p = 0$ when $\alpha < a_{\min}$ (Fig. 2B, top versus bottom panels).

Appendix 3 - Linked variation

Long-term balanced polymorphisms can affect patterns of genetic diversity at linked and otherwise neutrally evolving loci, leading to characteristic signatures of inflated heterozygosity in genomic regions where the balanced polymorphism resides. With the exceptions of trans-species polymorphisms, it is these signatures that are used by current methods for the inference of balancing selection in genomic data (Bitarello et al., 2018; Cheng & DeGiorgio, 2020; Siewert & Voight, 2017). For a balanced polymorphism with alleles A_1 and A_2 that are stably maintained at equilibrium frequencies $\hat{p}$ and $1 - \hat{p}$ (respectively), the mean heterozygosity at a nearby neutral site is approximately:

$$\bar{H} = 4N_e\mu\left[1 + 2\hat{p}(1-\hat{p})\left(\frac{1}{2N_eR} + \frac{(1-2\hat{p})^2}{2 + 8N_eR\hat{p}(1-\hat{p})}\right)\right] \tag{13}$$

where R is the total recombinational distance between the neutral site and the site of the balanced polymorphism, and μ is the mutation rate at the neutral site (the approximation, which assumes that $R >> \mu$, is obtained using mean coalescent times from p. 103 of Rice 2004).

The sum in the square brackets of eq. (13) determines the degree of inflation of heterozygosity at the linked site relative to that of an unlinked locus, where the expected heterozygosity would be $\bar{H} = 4N\mu$. Recombination rate and the equilibrium frequency of the selected site also influence the width of the spike in neutral diversity that could be seen around a long-term balanced polymorphism. Let $R = nr$ represent the cumulative recombination rate between the selected site and a given neutral site, where n is the number of nucleotide sites separating the two and r is the recombination rate per site (*i.e.*, the probability of a recombination event between two directly adjacent nucleotide sites). Assuming that the polymorphism is intermediate ($\hat{p} = 0.5$), then the window around the selected site over which neutral diversity is elevated by at least two-fold (*i.e.*: where $\frac{\bar{H}}{4N\mu} > 2$) will be $L = \frac{1}{2N_e r}$ nucleotides, which implies that species with large effective population sizes and large per-nucleotide recombination rates should exhibit particularly narrow windows of inflated heterozygosity around loci maintaining long-term balanced polymorphisms.