PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Special feature review

Cite this article: Kelly JK. 2021 The promise and deceit of genomic selection component analyses. *Proc. R. Soc. B* **288**: 20211812. https://doi.org/10.1098/rspb.2021.1812

Received: 13 August 2021
Accepted: 30 September 2021

Subject Category:
Evolution

Subject Areas:
evolution, genomics, theoretical biology

Keywords:
Mimulus, *Drosophila*, selection component, genomics

Author for correspondence:
John K. Kelly
e-mail: jkk@ku.edu

Dedicated to Bill Beavis's landmark paper (Beavis WD. 1994 The power and deceit of QTL experiments: lessons from comparative QTL studies. In *Forty-ninth Annual Corn and Sorghum Industry Research Conference*, pp. 250–266, Washington, DC: American Seed Trade Association).

Special Feature: Wild Quantitative Genomics: the genomic basis of fitness variation in natural populations, edited by Susan Johnston, Nancy Chen, Emily Josephs.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5665552.

THE ROYAL SOCIETY PUBLISHING

The promise and deceit of genomic selection component analyses

John K. Kelly

Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, USA

JKK, 0000-0001-9480-1252

Selection component analyses (SCA) relate individual genotype to fitness components such as viability, fecundity and mating success. SCA are based on population genetic models and yield selection estimates directly in terms of predicted allele frequency change. This paper explores the statistical properties of gSCA: experiments that apply SCA to genome-wide scoring of SNPs in field sampled individuals. Computer simulations indicate that gSCA involving a few thousand genotyped samples can detect allele frequency changes of the magnitude that has been documented in field experiments on diverse taxa. To detect selection, imprecise genotyping from low-level sequencing of large samples of individuals provides much greater power than precise genotyping of smaller samples. The simulations also demonstrate the efficacy of 'haplotype matching', a method to combine information from a limited collection of whole genome sequence (the reference panel) with the much larger sample of field individuals that are measured for fitness. Pooled sequencing is demonstrated as another way to increase statistical power. Finally, I discuss the interpretation of selection estimates in relation to the Beavis effect, the overestimation of selection intensities at significant loci.

1. Introduction

Quantitative genetics is a science of measurable variables. For this reason, it has proven an essential tool for field evolutionary biologists because key parameters like additive genetic variances and selection gradients can be estimated from whole-organism measurements [1,2]. Quantitative genetic models directly predict changes in quantities of interest such as the mean of a population that is experiencing inbreeding or natural selection. In this context, model inputs such as the additive genetic variance can be treated as a 'black box' [3]: a complicated function of unmeasurable elements that can be estimated only as an aggregate quantity. However, the development of genomic techniques now affords a peek inside this black box at how individual polymorphisms contribute to variation. Here, I will consider selection component analyses (SCA) [4–8] as one method to investigate the complicated relationship between individual loci and fitness.

SCA are statistical models that predict fitness components from observations of viability, fecundity and mating success. I will distinguish two SCA approaches: the '2-cohort' and 'family' designs. The 2-cohort design tests for differences in allele frequency between subdivisions (cohorts) of a population with distinct individuals constituting each cohort. Viability selection can be estimated from a contrast of individuals that survive to reproduce and those that do not [6,9]. Sexual selection can be measured by comparing individuals that acquire mates to those that do not [6,9–11]. The family design involves a contrast between non-independent relatives, specifically genotyped parents and their offspring [9,11]. Often only one parent can be genotyped: the one that is 'attached' to the progeny at sampling. This would be the female if sampling pregnant garter snakes (e.g. [12]), the male if sampling pregnant pipefish (e.g. [10]). When females are attached, the family design statistically distinguishes the observed maternal allelic contribution to offspring and from the inferred paternal contribution. We can then test whether the allele frequency

in reproductive females (p_A) differs from that in the population of *successful* male gametes (p_M). A significant difference between p_A and p_M indicates 'male selection' [9], which integrates a number of distinct selective mechanisms including sexual selection through differential siring and sperm/pollen competition.

SCA were initially applied to visible marker loci (e.g. [13]) or allozyme polymorphisms (e.g. [14]), but have recently been extended to test for selection at SNPs across the genome in a number of natural systems [10,11,15–20]. Genome-scale SCA, hereafter gSCA, provide a comprehensive evaluation of selection, but also confront difficult statistical challenges. The most basic is that per-locus effects will usually be small and thus hard to detect except in very large experiments. Genome-wide genotyping provides a great many opportunities to detect weak associations between genotype and fitness components, albeit with the cost that multiple testing corrections for thousands (or millions) of SNPs will make significance thresholds very stringent. The 'Beavis effect' [21] is an inevitable consequence: we underestimate the number of loci with fitness effects, but overestimate the effects of the subset of SNPs that fortuitously emerge as significant. One purpose of this paper is to examine the nature and magnitude of this effect for different gSCA experimental designs.

A second challenge for gSCA is that current sequencing technologies will often yield incomplete data. Many sampled individuals will have uncertain genotype calls at many or even most SNPs. gSCA are likelihood models and thus can naturally accommodate uncertain genotype calls simply by summing over all possible genotypes at a locus, weighted by their respective probabilities. In fact, the capacity of gSCA to accommodate fragmentary or incomplete data suggests a number of ways that researchers can increase the power of field studies. I will consider a number of these options in this paper, including low-level sequencing and pooled sequencing, by which experiments can include a much larger number of individuals without increased cost. Low-level sequencing on larger field collections (high uncertainty at the individual level) proves to be far more powerful than perfect genotyping of smaller datasets. With limited sequencing coverage, inference of selection on a particular SNP can be improved by considering sequence information from neighbouring SNPs. Capitalizing on localized LD is key to imputation methods [22,23] and also for the recently developed 'haplotype matching' approach for genomic gSCA [9].

Below, I first describe the structure of gSCA models and then use simulations to explore their statistical properties under a range of circumstances. While most cases are general, the examination of haplotype matching is grounded in genomic data from two species, *Mimulus guttatus* and *Drosophila melanogaster*. Haplotype matching was demonstrated in *M. guttatus*, a species with intermediate levels of linkage disequilibrium (LD); substantial within genes (inter-SNP distances of up to 1 kb) but declining to background levels between genes (approx. 20 kb) [24]. This is lower than in some species (e.g. *Arabidopsis* [25]) but higher than others (e.g. *Drosophila* [26]), and it is unclear how well haplotype matching will work when LD decays rapidly with distance between sites. The *Mimulus*/*Drosophila* contrast is also informative because while both species are amenable to gSCA, different sampling designs may be required. *Mimulus guttatus* is monoecious and thus the family design—a sample of maternal plants and their progeny—provides a direct contrast of allele frequencies in the population of successful male gametes (those contributing to progeny) with the overall adult male population. This is because the adult male population is the adult female population and thus maternal genotyping estimates both. By contrast, a distinct sampling of adult males, adult (pregnant) females and their progeny would be required in *Drosophila*. This three-sample experiment enables additional contrasts (e.g. a difference between adult males and females suggests sex-specific viability selection [4,20]) and basically combines the 2-cohort and family designs.

2. Theory and simulation

As an example of the 2-cohort test, consider a sample of n_A individuals that survive to adulthood and n_L individuals that do not. Each individual provides some genetic information about a particular bi-allelic SNP with alleles R and A. The log-transformed likelihood for the entire dataset can be written as

$$\mathrm{LnL} = \sum_{y=1}^{n_A} \mathrm{Ln}\left\{\sum_{i=0}^{2} X_A[G_y = i]P[\mathrm{Data}_y|G_y = i]\right\} + \sum_{z=1}^{n_L} \mathrm{Ln}\left\{\sum_{i=0}^{2} X_L[G_z = i]P[\mathrm{Data}_z|G_z = i]\right\}, \quad (2.1)$$

where $X[*]$ are the genotype frequencies among survivors and dead. Let p = frequency of the reference base (R). With random mating but no selection, $X_A[G_y = i] = X_L[G_y = i]$ with the values of p^2, $2p(1-p)$ and $(1-p)^2$, for $i = 0$, 1 and 2, respectively. Here, $i = 0$ for the reference base homozygote (RR), $i = 1$ for the heterozygote (RA) and $i = 2$ for the alternative homozygote (AA). With differences in viability, $X_A[G_y = i]$ and $X_L[G_y = i]$ can be written in terms of zygotic (pre-selection) allele frequency and genotype specific viabilities or selection coefficients. The nature of genomic data for an individual determines $P[\mathrm{Data}_y|G_y = i]$, the likelihood of data from individual y given that it has genotype $G_y = i$. With 'perfect genotyping' (no uncertainty), $P[\mathrm{Data}_y|G_y = i]$ is 1.0 for i matching the true G_y and zero otherwise. In this case, equation (2.1) simplifies to the multinomial likelihood of the original selection component models [4].

For a family design with a sample of n_F families, each with a maternal individual and n_O offspring, the log-likelihood for the dataset is

$$\mathrm{LnL} = \sum_{y=1}^{n_F} \mathrm{Ln}\left\{\sum_{i=0}^{2} X_A[G_y = i]P[\mathrm{Data}_y|G_y = i]\prod_{k=1}^{n_O} P[\mathrm{Data}_{yk}|G_y = i]\right\} \quad (2.2)$$

Here, $P[\mathrm{Data}_{yk}|G_y = i]$ is the probability of the data from offspring k of maternal individual y given that the maternal genotype is i. The product in equation (2.2) is a function of p_M, allele frequency in *successful* male gametes, while the $X_A[G_y = i]$ are a function of p_A, allele frequency in reproductive females. For the simulations, I assume independent siring of each offspring, but the model can be modified to accommodate any mixture of half and full siblings [11].

After considering cases with perfect genotyping, I limit the number of sequence reads that cover the focal SNP in individuals. With a single sequence read for the focal SNP, the individual will report either R or A, and the number of

possibilities for $P[\text{Data}_y|G_y = i]$ is limited. For $i = [0,1,2]$ in sequence,

$$P[\text{Data}_y|G_y = i] = [1.0 - \varepsilon, 0.5, \varepsilon] \quad \text{if the datum is R, and}$$
$$P[\text{Data}_y|G_y = i] = [\varepsilon, 0.5, 1.0 - \varepsilon] \quad \text{if the datum is A.}$$

Here, ε is the probability of the unlikely event that a homozygote reports the alternative allele owing to sequencing or bioinformatic error. With a single read, heterozygotes will always produce data that is most likely to have come from one of the homozygotes.

Pooled sequencing of progeny sets can greatly reduce the cost of family design experiments. For simulation of these cases, I assume that maternal genotype is determined without error: $G_y = 0$, 1 or 2. The data for a family are then G_y, m_R and m_A, where the latter terms refer to the number of sequence reads from the progeny pool that are R versus A at the focal SNP (ignoring sequencing error for these counts as well). n_O is the number of offspring in the pool. Given independent siring of offspring, the likelihood can be calculated simply by conditioning on the number of A alleles contributed by sires (0 up to n_O) to the entire progeny set. If $G_y = 0$, the likelihood for family y is

$$p_A^2 \sum_{i=0}^{n_O} \binom{n_O}{i} p_M^i (1-p_M)^{n_O - i} \binom{m_R + m_A}{m_R} \times \left(\frac{1}{2} + \frac{i}{2n}\right)^{m_R} \left(\frac{1}{2} - \frac{i}{2n}\right)^{m_A}. \tag{2.3}$$

If $G_y = 1$, the likelihood is

$$2p_A(1-p_A) \sum_{i=0}^{n_O} \binom{n_O}{i} p_M^i (1-p_M)^{n_O - i} \binom{m_R + m_A}{m_R} \times \left(\frac{1}{4} + \frac{i}{2n}\right)^{m_R} \left(\frac{3}{4} - \frac{i}{2n}\right)^{m_A}. \tag{2.4}$$

For $G_y = 2$,

$$(1-p_A)^2 \sum_{i=0}^{n_O} \binom{n_O}{i} p_M^i (1-p_M)^{n_O - i} \binom{m_R + m_A}{m_R} \times \left(\frac{i}{2n}\right)^{m_R} \left(1 - \frac{i}{2n}\right)^{m_A}. \tag{2.5}$$

The LnL for the entire dataset is the sum of log-transformed family likelihoods.

Haplotype matching requires a collection of high-quality whole-genome sequences from the natural population under study. In our prior experiment on *M. guttatus* [9], this 'reference set' consisted of 187 genomes. The field experiment involves sampling individuals measured for fitness, each of which is subject to low-level or incomplete sequencing (e.g. RADseq [27]). Instead of directly applying the gSCA models to the fragmentary data on field individuals, we align sequence reads (or read-pairs) to the reference set, treating the reference set as the haplotypes that segregate in the natural population. Given that read-pairs are routinely 200–300 bp in length, the data units from field individuals are themselves small haplotypes. In *M. guttatus*, read-pairs routinely overlap 5–10 SNPs. Consequently, a single read-pair can eliminate a large fraction of the reference panel as potential ancestors. Of course, there will be a great many more distinct haplotypes in the natural population than in the reference set if one considers long sequences. Thus, haplotype matching is conducted within delimited intervals (e.g. genes). In *Mimulus*, we found that field data obtained using MSG RADseq [28] were almost entirely congruent with the reference set within genes. In other words, there were few cases where field plants had sequence data that could not be reiterated by copying the sequence from at least one pair of reference set sequences. In the rare cases of inconsistency, field genotypes were treated as missing data. However, the high rate of matching enabled an effective probabilistic inference of SNP genotype associations with fitness. As a contrast to *Mimulus*, I here simulate comparable data from *D. melanogaster* using the Drosophila Genetic Reference Panel (DGRP) as the reference panel [29]. The DGRP is a set of 205 whole-genome sequences from a single natural population and is intended to represent a random sample of alleles from that population. Simulating low-level sequencing of flies from this population tests the feasibility of haplotype matching for organisms with lower intra-genic LD.

In haplotype matching, sequenced haplotypes from the reference set are treated as the alleles segregating in nature:

$$P[\text{Data}_y|G_y = k] = \frac{\sum_{i,j \geq i}^{\text{Ref lines}} P[M_y = i,j] P[\text{Data}_y|M_y = i,j] \delta_{ij,k}}{P[G = k]}. \tag{2.6}$$

Here, $\delta_{ij,k}$ is 1 if genic genotype $[ij]$ has SNP genotype k at the causal SNP and 0 otherwise. $P[M_y = i,j]$ is the probability that a random individual will carry genic genotype $[i,j]$, which is directly proportional to the frequency of these sequences in the reference panel. $P[G = k]$ is the Hardy–Weinberg frequency of genotype k (0, 1 or 2 at the focal SNP) within the reference panel. I denote each sequence in the set of reference lines as a 'genic haplotype' and diploid combinations as 'genic genotypes'. $P[\text{Data}_y|M_y = i,j]$ is the probability that all reads or read-pairs from this field individual would be produced by genic genotype $[i,j]$:

$$P[\text{Data}_y|M_y = i,j] = \prod_{r=1}^{\text{RP}} \left(\frac{B[h_{r,i}]}{2} + \frac{B[h_{r,j}]}{2}\right), \tag{2.7}$$

where RP is the number of read-pairs mapped within this gene for this individual, $h_{r,i}$ is the number of sequence mismatches between read-pair r and genic haplotype i, and $h_{r,j}$ is the corresponding value for haplotype j. $B[x]$ is the probability that a read-pair will exhibit $\times$ mismatches from the true sequence. For this simulation study, I will ignore sequencing errors and so $B[x] = 1$ for $x = 0$ and zero otherwise. However, applications to real data require an explicit model for these errors. It is natural that $B[x]$ should be proportional to ε^x, where ε is an error rate, although the proportionality might be adjusted to account for the number of SNPs overlapped by the read-pair.

Programs that implement equations (2.6) and (2.7) are published elsewhere [9] with modifications here to accept simulation data (programs in electronic supplementary material, file S1). For viability selection, we first simulate a sample of field individuals (survivors or dead) with data either at single SNPs or for entire genes. The sampling is contingent only on specified parameter values (e.g. p_A and p_L for SNP-specific simulations). The difference between p_A and p_L is varied to consider differing strengths of selection. For haplotype matching simulations, samples are taken from the full sequence sets of *M. guttatus* or *D. melanogaster*. The delineation of genic haplotypes for *M. guttatus* was done previously [9], and I here use a subset of that data for simulations. I chose A00309 (the median length gene on chromosome 1) as an

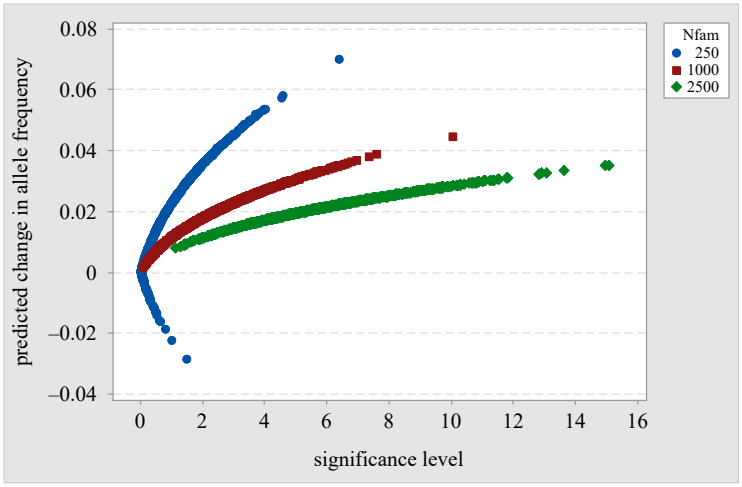

Figure 1. The predicted change in allele frequency at a SNP under male selection is given as a function of test significance level ($-\log_{10}(P)$) for 1000 simulation of each of three designs: the number of families (n_F) = 250 (blue), 1000 (red) or 2500 (green). Each family was genotyped for the maternal individual and five offspring. True $p = 0.5$ and true $\Delta p = 0.02$. (Online version in colour.)

'exemplar' for the first round of simulations and then randomly selected 100 genes from the remainder of chromosome 1 for the second round of simulations. For the SNP affecting fitness, I randomly selected from SNPs in a gene with a minor allele frequency greater than 0.1. The initial frequency, p, is determined by the frequency of bases in the sequenced lines. I elevate/reduce the frequencies of all reference lines that carry a base to simulate samples with different p_A and p_L. To create a simulated individual, I randomly select a pair of full sequences from the genic haplotype set for the gene. The probability of selecting a particular haplotype is proportional to its frequency in the reference panel (adjusted for selection). Given the two alleles within an individual, I simulate read-pairs as copies from these two sequences. For simplicity, the simulated read-pair is a continuous 300 bp copy of one of the two alleles. Each read-pair is initiated at a random position in the gene.

To generate data from the DGRP, I randomly sampled genes from chromosome 2 L with FBgn0010288 as the exemplar (it is the median length gene of 2 L: 1473 bp). I imputed missing values for each and extracted relevant data from each gene using the programs in electronic supplementary material, file S2. After processing, the genomic data from *Mimulus* and *Drosophila* are structurally equivalent as inputs to the same simulation/analysis programs. These programs estimate parameters, conduct likelihood ratio tests, and obtain *P*-values (*P* in upper case to distinguish from allele frequency) by comparing the likelihood ratio statistic (LRT) to chi-square-1. I calculated $-\log_{10}(P)$ as a measure of test significance from each simulated experiment. The mean of $-\log_{10}(P)$ across replicates is a measure of the power of the design to detect selection. The simulation programs for all cases are provided in electronic supplementary material, file S1. The programs original to this study were written in Python (v. 2.7). The 'Guide to programs' document in electronic supplementary material, file S1 provides instructions for use of these programs to generate results for any specified design. Finally, the figures presented below were made using Minitab v. 18.

3. Results

Both the promise and deceit of gSCA are illustrated by simulations without genotyping uncertainty. Figure 1 contrasts experiments of three different sizes applied to a population where the true $\Delta p = 0.02$ owing to male selection (family design). Estimation is unbiased for each design (the average predicted Δp is equal to the true value) and even the smallest experiment can detect selection. Marginal significance ($p < 0.05$) corresponds to $-\log_{10}(P) = 1.30$, a threshold exceeded by 31% of simulations with $n_F = 250$ families, 83% with $n_F = 1000$ and 99.8% for $n_F = 2500$. Critical for interpretation, however, is that there is a very strong relationship between estimates and test significance. For the small design ($n_F = 250$), the mean estimated Δp from *significant tests* (0.0356) is inflated by 78% relative to the true value. This bias is diminished with 1000 families and disappears entirely with 2500 families. In the latter case, there is no bias because nearly all simulations were significant at $p < 0.05$. However, genomic studies typically have stringent significance thresholds to correct for multiple testing. With $-\log_{10}(P) > 5$ as the threshold, about 60% of tests remain significant for the large design, but among these, Δp is exaggerated by an average of 30%. Similar trends are obtained with the 2-cohort test (electronic supplementary material, appendix A). A larger sweep of the parameter space for both family and 2-cohort tests is reported in electronic supplementary material, table S1.

Pooled sequencing of progeny is a potentially powerful tool for fecund organisms. Figure 2 contrasts perfect genotyping with pooled sequencing of progeny for the 'small' and 'medium' family designs ($n_F = 250$ or 1000). With 10–100 progeny per family, there is essentially no difference in power between perfect genotyping of progeny and 10× coverage of a DNA pool composed of all progeny. The latter option greatly reduces effort and cost relative to individual genotyping. Lower sequencing coverage of progeny pools (from 10× to 1× in figure 2) does reduce power, but only slightly. Note that 1× coverage of a pool with 100 progeny is superior to perfect genotyping of 10 offspring. With few offspring per family, pooling is less advantageous. The 2-cohort test is also amenable to pooled sequencing, and testing methods for this design are well established [30,31].

An important qualification on the results of figure 2 concerns the assumption that each offspring is sired independently. This is clearly inconsistent with monogamy, and most mating systems will produce a mixture of half and full sibs within maternal families. With individual genotyping of

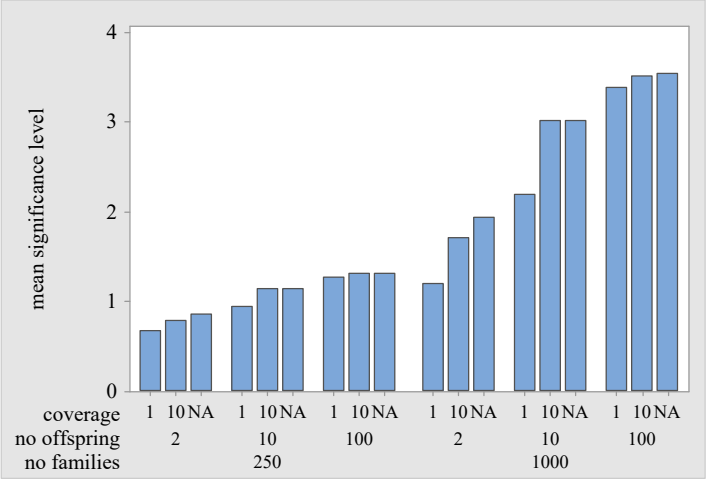

Figure 2. The mean significance level is reported for experiments varying in the number of families, the number of offspring per family and the sequencing coverage of each family. For the latter, NA refers to perfect individual genotyping of each offspring, 1 is 1 × coverage of a pooled DNA sample made from all offspring, and 10 is 10 × coverage of the same. The selection regime is the same as for figure 1 ($p = 0.5$, $\Delta p = 0.02$). (Online version in colour.)

progeny, it is straightforward to diagnose the 'internal' structure of maternal families (which offspring share a sire) from genome-wide SNP data [32,33]. Given this partitioning, the gSCA likelihood statements can be modified to account for the non-independence of full-siblings (see equation (2.4) in [11]), but this may not be possible with pooled progeny data. In this case, testing can fall back on the fact that each maternal family provides an unbiased and independent estimate for p_M, thus providing means for inference based on the average across families.

To evaluate low-level individual sequencing and haplotype matching, I focus on viability selection, although these techniques are fully compatible with the family design [9]. For these simulations considering SNPs across the genome of *Mimulus* or *Drosophila*, differing outcomes are generated by differences in initial allele frequency and differing patterns of local LD. To evaluate haplotype matching relative to single SNP genotyping, I simulated data based on the 'median' gene of *Mimulus* chromosome 1 (A00309, figure 3*a*). With the true $\Delta p = 0.01$ and SNP-specific data, power is much greater if one replaces perfect genotyping of 1000 individuals with low-coverage sequencing of 10 000 individuals (test 1 versus test 2 in figure 3*a*). If we consider tests at SNP 1566059, only 16% of test 1 cases are marginally significant ($p < 0.05$) in contrast to 51% for test 2. Across SNPs in figure 3*a*, 24% of test 1 are significant as opposed to 76% for test 2. For A00309, power is much higher with low-level sequencing (1× coverage) of the entire gene (haplotype matching) than with data specific to the selected site. Tests 2 and 3 each have an average of 1× coverage of the selected SNP, but the latter is much more likely to reveal selection (figure 3*a*). This is noteworthy given that over a third of test 3 cases have no coverage of the selected site. Leveraging haplotype structure greatly increases power to infer the genotype at the selected site. In fact, 1× coverage of A00309 is essentially indistinguishable from perfect genotyping in terms of power (case 3 versus case 4). For each of the 1*x* simulated datasets, I created a replicate dataset with the actual genotype output for model fit with no uncertainty. The correlation of LRT values between 1× and perfect genotyping is nearly 1 with the mean of the former value greater than 99% of latter.

Figure 3*b* considers the same strength of selection (true $\Delta p = 0.01$) for FBgn0010288, the median *Drosophila* gene. Simulations estimate power for gene-level data with 1 × coverage, 5 × coverage and perfect genotyping of 10 000 individuals. Comparing equivalent cases in figure 3*a*,*b*, the power of haplotype matching is lower with *Drosophila*. 1 × coverage is consistently and substantially weaker than perfect genotyping in the four tested SNPs. This is expected given lower LD in *Drosophila* relative to *Mimulus*. However, we find that 5 × coverage is nearly as good as perfect genotyping in three of four SNPs. The different outcome for the first SNP (2212741) illustrates that the effectiveness of haplotype matching depends on localized LD patterns, which can vary among SNPs. This variability is considered in greater detail below, but simulations across many genes demonstrate that figure 3 is representative.

I next performed simulations on 100 randomly selected genes from each species. We dropped one gene from each set owing to lack of variation. Within each remaining gene, we performed five replicate simulations each using a randomly selected SNP to determine fitness (minimum minor allele frequency of 0.1). The *Mimulus* results from A00309 (figure 3*a*) are largely reiterated by the broader scan (figure 4*a*). The power with 1 × coverage is only slightly lower than perfect genotyping of 10 000 individuals (case 3 versus case 4 in figure 4*a*). The 100 gene scan for *Drosophila* is also quite similar to the exemplar case (compare figure 3*b* to figure 4*b*), with 5 × coverage nearly indistinguishable from perfect genotyping.

To this point, simulations have sampled genic haplotypes from the set of referenced lines (187 in *Mimulus*, 205 in *Drosophila*). In real applications, it is likely that the reference sequence set will be incomplete. In other words, field individuals will have haplotypes that are not present in the reference set. To evaluate the consequences of this partial sampling, I conducted simulations using the entire reference sequences sets to generate field data, but fit the gSCA using only a random subset of 100 sequences as the 'experimental level' reference set. As consequence, simulated read-pairs may fail to map to any genic haplotype under consideration. Such individuals will have very low likelihoods for all the possible genotypes at a SNP and are treated as missing data in model fitting. I first applied the partial sampling pipeline to simulations where there is no selection and established that incompleteness of the reference panel does not inflate the

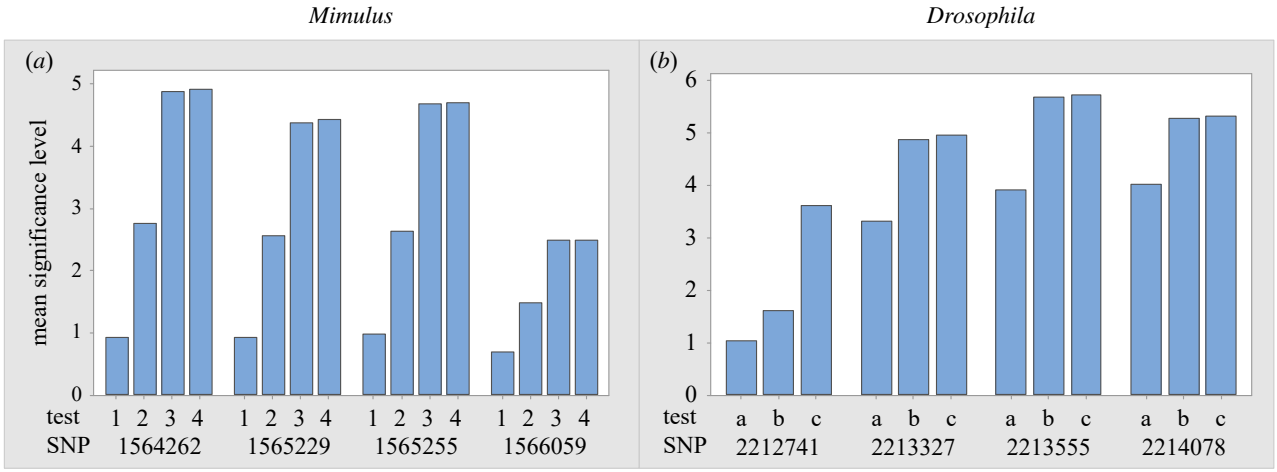

Figure 3. The mean significance level of tests for viability selection in simulations with the true $\Delta p = 0.01$. (*a*) Four SNPs within gene A00309 of *Mimulus*. Tests: 1 = perfect genotyping with $n_A = n_L = 500$, 2 = one read for causal SNP with $n_A = n_L = 5000$, 3 = 1 × coverage of gene coupled with haplotype matching with $n_A = n_L = 5000$, 4 = perfect genotyping with $n_A = n_L = 5000$. (*b*) Four SNPs within gene FBgn0010288 of *Drosophila*. All tests with $n_A = n_L = 5000$: *a* = 1 × genic with haplotype matching, *b* = 5 × genic with haplotype matching, *c* = perfect genotyping. (Online version in colour.)

probability of false positives. Considering cases with selection, incomplete reference panels can reduce power, but the effect is surprisingly weak (figure 4*c*,*d*). For the 99 gene set in *Mimulus*, 1 × sequencing is only slightly different when mapping reads to the incomplete reference set as opposed to the full set (figure 4*c*). In *Drosophila*, we performed contrasts with both 1 × and 5 × coverage. In both cases, mapping to the partial set reduced power, but only slightly (figure 4*d*).

4. Discussion

Sample size is the fundamental challenge in the direct study of natural selection. Experiments that relate the genotype of individual organisms to observed survival and reproduction provide an immediate view of natural selection. Unfortunately, precision is a limiting factor for field experiments. Small fitness differences (less than or equal to 1%) may be sufficient for natural selection to overwhelm other forces, but measurements on thousands of individuals are required to accurately estimate such effects [34]. As an alternative to direct study, many evolutionary biologists search for natural selection through statistical analyses of sequence polymorphism and divergence patterns [35–37]. This molecular approach has the advantage of integrating small allele frequency changes that accrue over many generations and these procedures have identified selected loci across a range of species [38–40]. Regrettably, sequence variation patterns provide limited information about the fitness differences between genotypes. Different selective mechanisms can produce indistinguishable patterns in sequence data [41] and variable selection (an inconstant genotype-to-fitness mapping) can generate patterns that mimic neutrality. Inconstancy can result from temporal fluctuations in selection [42,43], or frequency dependent fitnesses [44], or even simple quantitative inheritance where the phenotype (and not the genotype) is the immediate effector of fitness [45]. Given these difficulties, the direct estimation of natural selection within wild populations remains essential.

Genomic selection component analyses (gSCA) assess SNP-fitness associations across the genome. The simulations illustrated by figures 1–4 show that gSCA involving a few thousand genotyped samples can detect selection if changes in allele frequency (Δp) are 1% or greater. This is noteworthy given that field surveys suggest per generation Δp can be greater than 1% at hundreds of SNPs in plants [9,11,17], insects [15,43,46] and even vertebrates [10,19]. Regarding experimental design, the primary message from this simulation study is that researchers should endeavour to include as many genomes as possible in an experiment. Greater biological replication is strongly favoured at the expense of precision in individual genotyping. Low-level sequencing is effective, particularly if aided by haplotype matching (figures 3 and 4). Pooled sequencing can be nearly as powerful as individual sequencing (figure 2) allowing researchers to sequence thousands of individuals without having to make thousands of sequencing libraries. The pooled-progeny option is particularly attractive for highly fecund organisms, such as plants and many invertebrates, where hundreds or even thousands of progeny can be sampled from a single maternal individual.

Haplotype matching with low-level sequencing proved more effective in *Mimulus* than *Drosophila* (figures 3 and 4), a result with at least two causes. First, nucleotide diversity is substantially higher in the Iron Mountain population of *M. guttatus* (source of the *Mimulus* lines) than in the DGRP of *Drosophila* [24,47]. *Mimulus* read-pairs will thus overlap more SNPs and more incisively identify ancestral haplotypes. Second, haplotype structure (localized LD) is higher in *Mimulus* than *Drosophila*. This facilitates the 'process of elimination' aspect of haplotype matching simply because fewer genotypic combinations are present (or abundant) in ancestors. While 1*x* haplotype matching is not as good in *Drosophila* as *Mimulus*, the procedure works surprisingly well for a species where LD is often considered negligible. Indeed, averages for pairwise measures of LD, such as r^2, decline rapidly with distance in the DGRP [47]. However, r^2 is an imperfect predictor of the number of distinct sequences that exist for a gene. For the 99 genes sampled from chromosome 2 L (figure 4), the median number of distinct sequences (per gene) was only 82. This is far fewer than would be expected among 205 sequences if SNP alleles were sampled independently along each gene (given the number of SNPs and allele frequencies).

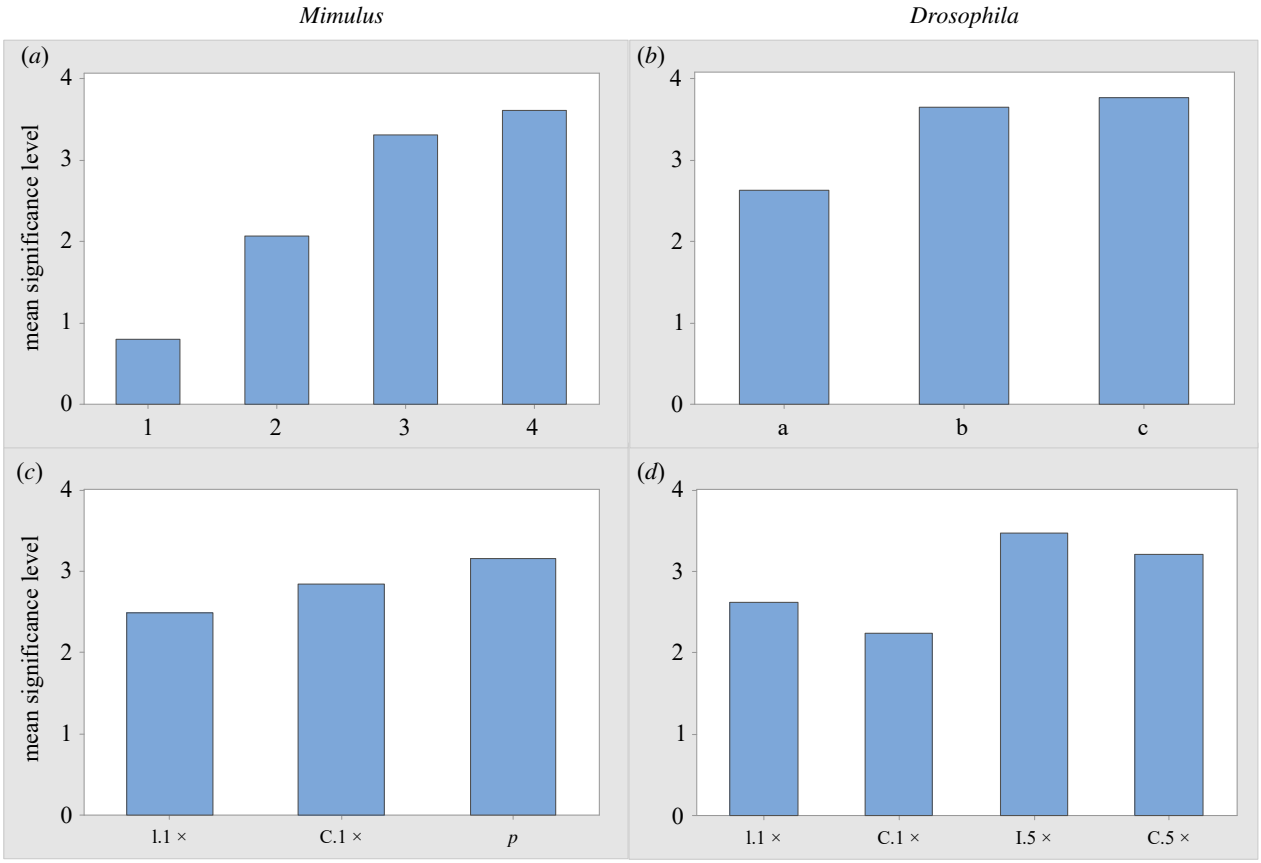

Figure 4. Mean significance level is reported for viability selection across 99 randomly selected genes of *Mimulus* (*a*,*c*) or *Drosophila* (*b*,*d*). For (*a*) and (*b*), the selection regime and tests match their corresponding IDs in figure 3*a*,*b*. For (*c*) and (*d*), true $\Delta p = 0.03$ with 1000 individuals were measured for all cases. For (*c*), I.1 × = incomplete reference set with 1 × sequencing coverage (genic), C.1 × = complete reference set with 1 × coverage and *p* = perfect genotyping. For (*d*), *I*/*C* is incomplete versus complete reference set and 1 ×/5 × is coverage. (Online version in colour.)

All simulated cases display a difficulty with gSCA, the tendency to overestimate the strength of selection at SNPs that emerge as significant (figure 1). For any given scenario (selection strength and experimental design), the most significant tests are those with the most fortuitous estimation error (i.e. those that most exaggerate Δp). Across different scenarios, we find that as the power of a test declines, the bias increases among significant estimates. The biases among ascertained estimates from gSCA are merely the most recent reiteration of the 'Beavis effect' [21]. Described originally for QTL mapping experiments, this tendency to overestimate effects emerges routinely in big data analyses [48] where investigators perform large numbers of tests, rank them by significance and claim the most extreme outcomes as positive results.

(a) gSCA versus GWAS

Genome-wide association studies estimate SNP-specific effects on quantitative traits and can be applied to fitness components. In some cases, gSCA and genome-wide association (GWAS) models can be applied to the same experiment. Consider the 2-cohort design with a genotyped population sorted into survivors and deceased. The gSCA on these data tests for allele or genotype frequency divergence between cohorts. The relevant GWAS is a generalized linear model with genotype as the factor and survival as a binary response variable [49]. gSCA works in the currency of allele frequency change (Δp) while GWAS in terms of genotypic effects, but with equivalent assumptions, parameters should be inter-convertible. However, the differing objectives of gSCA and GWAS are important.

GWAS models typically include a random effect for genetic background to absorb the diffuse associations between the focal SNP and loci across the genome. Such associations can be caused by unrecognized population structure or admixture. With gSCA, the focus is prediction of Δp owing either to selection acting directly on the SNP or via hitchhiking [50]. Indeed, the effectiveness of haplotype matching depends on local LD, but such associations make it very difficult to distinguish causal SNPs from hitch-hikers. Second, GWAS but not gSCA routinely include environmental factors as covariates to statistically remove their effects [51]. gSCA models can be written to include covariates, but with a focus on natural populations, one may not wish to factor out environmental effects. Consider a locus that affects habitat choice. The action of this locus will generate a genotype by environment correlation. If habitat subsequently affects fitness, allele frequency change will result. gSCA will detect this Δp despite the absence of a genotype/fitness association *within* habitats. Of course, there are situations where gSCA and GWAS models can be applied jointly. Consider a sampling of maternal plants and their seedset. gSCA can detect male selection through the contrast of allele frequencies between mothers and progeny. GWAS models can estimate female fecundity selection by predicting seed number from the individual genotype of maternal plants.

Estimates from gSCA and GWAS are equally susceptible to the Beavis effect. One antidote is to repeat the association study in a distinct and independent panel. In GWAS of medically relevant traits of humans, replication across large panels is taken as evidence of genuine effect [52]. Two things make

this option less feasible for gSCA or GWAS of fitness components in wild populations. First, cost is a serious difficulty given the more limited resources available for natural population studies. Second, it may simply be impossible to find replicate panels for fitness. For most species, the mapping from genotype to fitness will vary from one population to the next, and from one generation to the next within a population [53]. In this situation, averaging across 'replicates' cancels signal as well as noise.

Given spatial and temporal variation in selection, corroboration for gSCA estimates must come in another form. Our focal species, *D. melanogaster* and *M. guttatus*, provide a few examples. In *D. melanogaster*, large amplitude fluctuations in allele frequency occur seasonally and Δp can be predicted from weather conditions [54]. Δp generated by estimation error or population genetic drift will not correlate with environmental variables. In *M. guttatus*, SNPs under viability selection within one population exhibit elevated allele frequency divergence from other populations relative to 'neutral' SNPs, and the direction of selection (the sign of Δp) predicts the direction of divergence [11]. In both *Mimulus* and *Drosophila*, SNPs currently under selection in field populations exhibit molecular population genetic signatures of long-term selection [9,46], which is a distinct sort of corroboration. These are useful examples, but perhaps the most incisive contrasts have yet to be developed as field biologists increasingly apply genomic methods to the study of natural selection in wild populations.

Data accessibility. The data are provided in the electronic supplementary material [55].

Competing interests. I declare I have no competing interests.

Funding. I received support from National Institutes of Health grant no. R01 GM073990 and NSF grant no. 1753630.

Acknowledgements. I thank S. Macdonald, L. Fishman, P. Monnahan, R. Unckless, B. Servin, S. Johnston and an anonymous referee for advice and/or edits to the manuscript. This work was supported by the HPC facilities operated by the Center for Research Computing at the University of Kansas.

References

1. Arnold SJ. 1994 Multivariate inheritance and evolution: a review of concepts. In *Quantitative genetic studies of behavioral evolution* (ed. CRB Boake), pp. 17–48. Chicago, IL: University of Chicago Press.
2. Grant PR, Grant BR. 1995 Predicting microevolutionary responses to directional selection on heritable variation. *Evolution* **49**, 241–251. (doi:10.1111/j.1558-5646.1995.tb02236.x)
3. Hill WG. 2010 Understanding and using quantitative genetic variation. *Phil. Trans. R. Soc. B* **365**, 73–85. (doi:10.1098/rstb.2009.0203)
4. Christiansen F, Frydenberg O. 1973 Selection component analysis of natural polymorphisms using population samples including mother-offspring combinations. *Theor. Popul. Biol.* **4**, 425–445. (doi:10.1016/0040-5809(73)90019-1)
5. Allard RW, Kahler AL, Clegg MT. 1977 Estimation of mating cycle components of selection in plants. In *Measuring selection in natural populations* (eds FB Christiansen, TM Fenchel). Berlin, Heidelberg: Springer.
6. Bundgaard J., Christiansen FB. 1972 Dynamics of polymorphisms. I. Selection components in an experimental population of *Drosophila melanogaster*. *Genetics* **71**, 439–460. (doi:10.1093/genetics/71.3.439)
7. Prout T. 1965 The estimation of fitness from genotypic frequencies. *Evolution* **19**, 546–551. (doi:10.1111/j.1558-5646.1965.tb03330.x)
8. Knight GR, Robertson A. 1957 Fitness as a measurable character in *Drosophila*. *Genetics* **42**, 524–530. (doi:10.1093/genetics/42.4.524)
9. Monnahan PJ, Colicchio J., Fishman L., Macdonald SJ, Kelly JK. 2021 Predicting evolutionary change at the DNA level in a natural *Mimulus* population. *PLoS Genet.* **17**, e1008945. (doi:10.1371/journal.pgen.1008945)
10. Flanagan SP, Jones AG. 2017 Genome-wide selection components analysis in a fish with male pregnancy. *Evolution* **71**, 1096–1105. (doi:10.1111/evo.13173)
11. Monnahan PJ, Colicchio J., Kelly JK. 2015 A genomic selection component analysis characterizes migration-selection balance. *Evolution* **69**, 1713–1727. (doi:10.1111/evo.12698)
12. Arnold SJ. 1981 Behavioral variation in natural populations. I. phenotypic, genetic and environmental correlations between chemoreceptive responses to prey in the garter snake, *Thamnophis elegans*. *Evolution* **35**, 489–509. (doi:10.1111/j.1558-5646.1981.tb04912.x)
13. Prout T. 1971 The relation between fitness components and population prediction in *Drosophila*. I. The estimation of fitness components. *Genetics* **68**, 127–149. (doi:10.1093/genetics/68.1.127)
14. Clegg MT, Kahler AL, Allard RW. 1978 Estimation of life cycle components of selection in an experimental plant population. *Genetics* **89**, 765–792. (doi:10.1093/genetics/89.4.765)
15. Soria-Carrasco V *et al*. 2014 Stick insect genomes reveal natural selection's role in parallel speciation. *Science* **344**, 738–742. (doi:10.1126/science.1252136)
16. Anderson JT, Lee C-R, Mitchell-Olds T. 2014 Strong selection genome-wide enhances fitness trade-offs across environments and episodes of selection. *Evolution* **68**, 16–31. (doi:10.1111/evo.12259)
17. Troth A, Puzey JR, Kim RS, Willis JH, Kelly JK. 2018 Selective trade-offs maintain alleles underpinning complex trait variation in plants. *Science* **361**, 475–478. (doi:10.1126/science.aat5760)
18. Exposito-Alonso M *et al*. 2019 Natural selection on the *Arabidopsis thaliana* genome in present and future climates. *Nature* **573**, 126–129. (doi:10.1038/s41586-019-1520-9)
19. Chen N, Juric I, Cosgrove EJ, Bowman R, Fitzpatrick JW, Schoech SJ, Clark AG, Coop G. 2019 Allele frequency dynamics in a pedigreed natural population. *Proc. Natl Acad. Sci. USA* **116**, 2158–2164. (doi:10.1073/pnas.1813852116)
20. Cheng C, Kirkpatrick M. 2016 Sex-specific selection and sex-biased gene expression in humans and flies. *PLoS Genet.* **12**, e1006170. (doi:10.1371/journal.pgen.1006170).
21. Beavis WD. 1994 The power and deceit of QTL experiments: lessons from comparative QTL studies. In *Forty-ninth Annual Corn and Sorghum Industry Research Conference, Chicago, IL, 7–8 December*, pp. 250–266. Washington, DC: American Seed Trade Association.
22. Shi S, Yuan N, Yang M, Du Z, Wang J, Sheng X, Wu J, Xiao J. 2018 Comprehensive assessment of genotype imputation performance. *Hum. Hered.* **83**, 107–116. (doi:10.1159/000489758)
23. Davies RW, Flint J, Myers S, Mott R. 2016 Rapid genotype imputation from sequence without reference panels. *Nat. Genet.* **48**, 965–969. (doi:10.1038/ng.3594)
24. Puzey JR, Willis JH, Kelly JK. 2017 Population structure and local selection yield high genomic variation in *Mimulus guttatus*. *Mol. Ecol.* **26**, 519–535. (doi:10.1111/mec.13922)
25. Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M. 2007 Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **39**, 1151–1155. (doi:10.1038/ng2115)
26. Houle D., Márquez EJ. 2015 Linkage disequilibrium and inversion-typing of the *Drosophila melanogaster* genome reference panel. *G3* **5**, 1695–1701. (doi:10.1534/g3.115.019554)
27. Davey JW, Blaxter ML. 2010 RADSeq: next-generation population genetics. *Brief. Funct. Genom.* **9**, 416–423. (doi:10.1093/bfgp/elq031)

28. Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T, Stern DL. 2011 Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.* **21**, 610–617. (doi:10.1101/gr.115402.110)
29. Mackay TFC *et al*. 2012 The *Drosophila melanogaster* genetic reference panel. *Nature* **482**, 173–178. (doi:10.1038/nature10811)
30. Monnahan PJ, Kelly JK. 2017 The genomic architecture of flowering time varies across space and time in *Mimulus guttatus*. *Genetics* **206**, 1621–1635. (doi:10.1534/genetics.117.201483)
31. Vlachos C, Burny C, Pelizzola M, Borges R, Futschik A, Kofler R, Schlötterer C. 2019 Benchmarking software tools for detecting and quantifying selection in evolve and resequencing studies. *Genome Biol.* **20**, 169. (doi:10.1186/s13059-019-1770-8)
32. Ellis TJ, Field DL, Barton NH. 2018 Efficient inference of paternity and sibship inference given known maternity via hierarchical clustering. *Mol. Ecol. Resour.* **18**, 988–999. (doi:10.1111/1755-0998.12782)
33. Gibson MJS, Crawford DJ, Holder MT, Mort ME, Kerbs B, de Sequeira MM, Kelly JK. 2020 Genome-wide genotyping estimates mating system parameters and paternity in the island species *Tolpis succulenta*. *Am. J. Bot.* **107**, 1189–1197. (doi:10.1002/ajb2.1515)
34. Lewontin RC. 1974 *The genetic basis of evolutionary change*. New York, NY: Columbia University Press.
35. Casillas S, Barbadilla A. 2017 Molecular population genetics. *Genetics* **205**, 1003–1035. (doi:10.1534/genetics.116.196493)
36. Nielsen R. 2005 Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218. (doi:10.1146/annurev.genet.39.073003.112420)
37. Cutter AD, Payseur BA. 2013 Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat. Rev. Genet.* **14**, 262–274. (doi:10.1038/nrg3425)
38. Hill T., Koseva BS, Unckless RL. 2019 The genome of *Drosophila innubila* reveals lineage-specific patterns of selection in immune genes. *Mol. Biol. Evol.* **36**, 1405–1417. (doi:10.1093/molbev/msz059)
39. Wright SI, Gaut BS. 2004 Molecular population genetics and the search for adaptive evolution in plants. *Mol. Biol. Evol.* **22**, 506–519. (doi:10.1093/molbev/msi035)
40. Filatov DA, Charlesworth D. 1999 DNA polymorphism, haplotype structure and balancing selection in the Leavenworthia PgiC Locus. *Genetics* **153**, 1423–1434. (doi:10.1093/genetics/153.3.1423)
41. Takahata N, Nei M. 1990 Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* **124**, 967–978. (doi:10.1093/genetics/124.4.967)
42. Gillespie JH. 1994 *The causes of molecular evolution*. Oxford, UK: Oxford University Press.
43. Wittmann MJ, Bergland AO, Feldman MW, Schmidt PS, Petrov DA. 2017 Seasonally fluctuating selection can maintain polymorphism at many loci via segregation lift. *Proc. Natl Acad. Sci. USA* **114**, E9932–E9941. (doi:10.1073/pnas.1702994114)
44. Ejsmond MJ, Babik W, Radwan J. 2010 MHC allele frequency distributions under parasite-driven selection: a simulation model. *BMC Evol. Biol.* **10**, 332. (doi:10.1186/1471-2148-10-332)
45. Kelly JK. 2006 Geographical variation in selection, from phenotypes to molecules. *Am. Nat.* **167**, 481–495. (doi:10.1086/501167)
46. Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA. 2014 Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. *PLoS Genet.* **10**, e1004775. (doi:10.1371/journal.pgen.1004775).
47. Huang W *et al*. 2014 Natural variation in genome architecture among 205 *Drosophila melanogaster* genetic reference panel lines. *Genome Res.* **24**, 1193–1208. (doi:10.1101/gr.171546.113)
48. Ioannidis J. 2008 Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (doi:10.1097/EDE.0b013e31818131e7)
49. Prentice RL, Pyke R. 1979 Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411. (doi:10.1093/biomet/66.3.403)
50. Maynard SJ, Haigh J. 1974 The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35. (doi:10.1017/S0016672300014634)
51. Wang T, Xue X, Xie X, Ye K, Zhu X, Elston RC. 2018 Adjustment for covariates using summary statistics of genome-wide association studies. *Genet. Epidemiol.* **42**, 812–825. (doi:10.1002/gepi.22148)
52. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22. (doi:10.1016/j.ajhg.2017.06.005)
53. Kingsolver JG, Hoekstra HE, Hoekstra JM, Berrigan D, Vignieri SN, Hill CE, Hoang A, Gibert P, Beerli P. 2001 The strength of phenotypic selection in natural populations. *Am. Nat.* **157**, 245–261. (doi:10.1086/319193)
54. Machado HE *et al*. 2019 Broad geographic sampling reveals predictable, pervasive, and strong seasonal adaptation in *Drosophila*. *bioRxiv*, 337543. (doi:10.1101/337543)
55. Kelly JK. 2021 The promise and deceit of genomic selection component analyses. Figshare.