Title

Controlling for *p*-value inflation in allele frequency change in experimental evolution and artificial selection experiments.

Authors

Petri Kemppainen*, Bernt Rønning*, Thomas Kvalnes*, Ingerid J. Hagen*, Thor-Harald Ringsby*, Anna M. Billing*, Henrik Pärn*, Sigbjørn Lien†, Arild Husby*‡, Bernt-Erik Sæther* and Henrik Jensen*

Affiliations

*Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway.

† CIGENE, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Ås, Norway.

‡ Department of Biosciences, P.O. Box 65 (Viikinkaari 1), 00014 University of Helsinki, Finland.

Keywords

Experimental evolution, artificial selection, population stratification, *p*-value inflation, relatedness, genome wide association studies

Corresponding author

Petri Kemppainen, Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, Høgskoleringen 5, Realfagbygget E1-126, NO-7491, Trondheim, Norway.

Phone number: +4745394509

E-mail address: petrikemppainen2@gmail.com

Running title

P-value inflation in experimental evolution

Abstract

Experimental evolution studies can be used to explore genomic response to artificial and natural selection. In such studies, loci that display larger allele frequency change than expected by genetic drift alone are assumed to be directly or indirectly associated with traits under selection. However, such studies report surprisingly many loci under selection, suggesting that current tests for allele frequency change may be subject to *p*-value inflation and hence be anti-conservative. One factor known from genome wide association (GWA) studies to cause *p*-value inflation is population stratification, such as relatedness among individuals. Here we suggest that by treating presence of an individual in a population after selection as a binary response variable, existing GWA methods can be used to account for relatedness when estimating allele frequency change. We show that accounting for relatedness like this effectively reduces false positives in tests for allele frequency change in simulated data with varying levels of population structure. However, once relatedness has been accounted for, the power to detect causal loci under selection is low. Finally, we demonstrate the presence of *p*-value inflation in allele frequency change in empirical data spanning multiple generations from an artificial selection experiment on tarsus length in two wild populations of house sparrow, and correct for this using genomic control. Our results indicate that since allele frequencies in large parts of the genome may change when selection acts on a heritable trait, such selection is likely to have considerable and immediate consequences for the eco-evolutionary dynamics of the affected populations.

Introduction

Phenotypic evolution experiments have been imperative for our understanding of both short and long-term evolutionary responses to selection (Dudley *et al.* 1977; Palmer & Dingle 1986; Gromko *et al.* 1991; Hill & Caballero 1992; Gromko 1995; Brakefield 2003; Conner 2003; Garland 2003). With increasing availability of population genomic data, it has become feasible to target the genomic changes that underlie phenotypic changes in such experiments (Ellegren & Sheldon 2008; Pardo-Diaz *et al.* 2015; Schlötterer *et al.* 2015). Two approaches that can be used to study genomic responses of selection are; (1) artificial selection, where individual survival or reproduction is artificially manipulated based on traits of interest (Heidaritabar *et al.* 2014) and (2) natural selection experiments, where survival and reproduction instead depends on the individuals inherent ability to cope with the environmental conditions (laboratory or natural) they are subjected to (Burke *et al.* 2010; Zhou *et al.* 2011; Turner *et al.* 2011; Remolina *et al.* 2012; Pespeni *et al.* 2013; Tobler *et al.* 2014; Gompert *et al.* 2014; Schlötterer *et al.* 2015). These studies often assume that loci showing significant allele frequency change following an episode of selection (e.g. when observed change falls outside the 95% quantiles of an appropriate null-distribution) are associated with the trait under selection (Barrett & Hoekstra 2011; Pespeni *et al.* 2013; Gompert *et al.* 2014; Heidaritabar *et al.* 2014). Such associations can stem from loci directly affecting the trait under selection, or indirectly through genetic correlations deriving from linkage disequilibrium (LD; Nielsen 2005; Barrett & Hoekstra 2011). Studies of allele frequency change following episodes of selection like this are valuable because they can give insights into both the

number and the type of genes associated with potentially highly complex adaptations.

Genome wide association (GWA) studies are powerful tools to dissect the genetic architecture of quantitative and binary traits (McCarthy *et al.* 2008; Bush & Moore 2012). In such studies, it is widely recognized that relatedness at any level of the population hierarchy, ranging from family structure to population structure at different spatial scales (here collectively referred to as population stratification) may cause long range LD between loci (Korte & Farlow 2013). In turn, this may lead to false association between genotypes and phenotypes, often evident as substantial *p*-value inflation and large numbers of false positives (Devlin & Roeder 1999; Devlin *et al.* 2001; Marchini *et al.* 2004; Price *et al.* 2010). As in GWA studies, test statistics for allele frequency change in experimental evolution rely on associations between genotypes and phenotypes. However, the possibility of *p*-value inflation due to population stratification in tests for allele frequency change have repeatedly been overlooked (Burke *et al.* 2010; Zhou *et al.* 2011; Turner *et al.* 2011; Turner & Miller 2012; Remolina *et al.* 2012; Pespeni *et al.* 2013; Turner *et al.* 2013; Gompert *et al.* 2014; Heidaritabar *et al.* 2014). These studies have consequently identified a surprisingly large number of loci putatively under selection (i.e. candidate loci). These findings were first questioned by Tobler *et al.* (2014), who showed that most of the identified candidate SNPs indeed were false positives, both by replicated experiments in *Drosophila melanogaster,* and in simulations. The false positives were mainly attributed to long range LD; either occurring naturally in the population (due to undetected population stratification) or as a consequence of the founders in the experiment representing only a small sample of the much

larger natural population. The mechanisms that cause *p*-value inflation in GWA studies are potentially the same that cause *p*-value inflation in allele frequency change in experimental evolution. While showing the potential for *p*-value inflation, Tobler *et al.* (2014) did not suggest any approaches to estimate its magnitude or to adjust for it. Here we demonstrate how methods already available to account for *p*-value inflation in GWA studies can be applied to genomic data from experimental evolution studies as well.

An appealing approach to study the effects of selection on genome variation is to estimate the population mean allele frequency change before and after selection (Pespeni *et al.* 2013; Gompert *et al.* 2014). If these episodes of selection occur within a single generation, the effects of drift and selection on such allele frequency change (estimated separately for each individual locus) are isolated from other processes, such as recombination and mutation, and empirical null-distributions can be generated by random permutation of samples (Pespeni *et al.* 2013; Gompert *et al.* 2014). As random permutation of samples does not take into account relatedness between individuals, we here demonstrate with simulations that estimating significance of allele frequency change like this is highly susceptible to *p*-value inflation arising from population stratification. As a means to account for *p*-value inflation, we propose that allele frequency change before and after selection can be tested using binary GWA analyses, where relatedness is included as a random effect (Aulchenko *et al.* 2007). Such tests are applicable for data sets where samples of individuals are individually genotyped prior to a single episode of natural or artificial selection, and the same individuals can be classified as either present or absent in the population following the selection episode. Hence, we have here not considered

131 other types of data such as those from pooled sequencing experiments (e.g. Parts
132 *et al.* 2011; Illingworth *et al.* 2012).

133 Whenever residual *p*-value inflation exists in the data, it is common
134 practice in GWA studies to perform genomic control (GC; Price *et al.* 2010). The
135 inflation factor (λ) can be estimated by regression in a Q-Q plot, comparing
136 observed versus expected (under the null-distribution) association statistics
137 (Clayton *et al.* 2005), and GC is subsequently achieved by dividing the observed
138 association statistics by λ. We test the merits of binary GWA analyses and GC on
139 allele frequency change before and after selection using simulated population
140 genomic data with varying levels of population structure. To demonstrate the
141 close relationship between testing for allele frequency change in a GWA
142 framework like this, and GWA analyses on the underlying quantitative trait
143 under selection, we also compare results from the two different approaches,
144 when relevant. The correlation between *p*-values from these two tests will give
145 an indication to what extent they identify the same genomic regions being
146 associated with the trait under selection.

147 Finally, as a demonstration of the concepts developed, we evaluate the
148 occurrence of *p*-value inflation on empirical SNP data from an artificial selection
149 experiment on two free-living island populations of house sparrow (*Passer*
150 *domesticus*). In the experiment, tarsus length was artificially selected to increase
151 or decrease across four consecutive years (2002-2005), resulting in an average
152 phenotypic change of 0.5-0.6% per year in the expected directions (Kvalnes et
153 al., in review). Furthermore, it was shown that this change had a genetic basis:
154 the average breeding values for tarsus length of cohorts produced on the two
155 islands during these four years also changed in the directions predicted by the

artificial selection, and these changes were larger than expected due to genetic drift (Kvalnes et al., in review). Due to overlapping generations in the house sparrow (Jensen et al. 2008), allele frequency change over the whole experimental period cannot easily be tested directly with binary GWA analyses. Instead, *p*-values for allele frequency change were obtained from empirical null-distributions produced by gene-dropping simulations and represents thus a more complex study design compared to estimating allele frequency change within a single generation.

Materials and Methods

Simulated population genomic data

Simulated population genomic data sets were generated with the software *fastsimcoal2* (Excoffier & Foll 2011; Excoffier *et al.* 2013) with three chromosomes of 1Mb each, mutation rate of = $3x10^{-8}$, recombination rate of $1x10^{-8}$ and no transition bias. With these parameters at least 5000 polymorphic SNPs were generated for all data sets. In data sets without population structure ('random mating') we set the effective population size (N_e) to 20000. This is equivalent to two populations of N_e=10000, each exchanging half of the population as migrants each generation (i.e. $N_em = N_e/2$, where m is the proportion of migrants exchanged each generation). In data sets with population structure we set the number of populations to two with N_e = 10000 each and N_em = 2 ('moderate population structure') or N_em = 1 ('strong population structure'). A relatively large N_e ensured that LD quickly declines with physical distance. From simulations with no population structure we sampled 100 diploid individuals and with population structure we sampled 50 diploid individuals

from each of the two populations. Five thousand bi-allelic SNPs with a minor allele frequency (MAF) above 0.05 were randomly chosen to create data sets of equal sizes for all levels of population structure.

For each replicate simulated data set, two, four or eight loci were randomly chosen to represent causal loci. For each causal locus, one allele was randomly chosen to translate to a phenotypic value of one with the alternative translating to a phenotypic value of zero, giving phenotypic values of 0, 1 or 2 for genotypes at each causal locus. The final phenotypic value for each individual was the sum of these values across the causal loci, with Gaussian noise added to generate a narrow sense heritability of $h^2 = 0.5$, defined as V_A/V_P, where V_P is the total phenotypic variance and V_A the additive genetic variance (which were known in our simulated data). Individuals with phenotypic values above the mean plus 0.3 standard deviations of the mean were considered as 'surviving' corresponding to an average selection intensity of one (Falconer & Mackay 1996). To simulate no heritability, phenotypic values were randomized among individuals prior to analyses. For each combination of levels in population structure and number of causal loci we generated 100 replicates, resulting in a total 900 of simulated data sets. In analyses with and without heritability the same simulated data sets were used.

Linkage disequilibrium

Linkage disequilibrium is well known to increase with population structure. Here we present analyses of LD of the simulated data sets mainly as a background for discussing its role in causing *p*-value inflation in tests for allele frequency change. Linkage disequilibrium was estimated as the coefficient of

determination between pairs of loci (r^2) for all pairwise comparisons between 500 randomly chosen SNPs from each simulated data set using the function *r2fast* from the R-package *GenABEL* (Hao *et al.* 2007; Aulchenko *et al.* 2007). Linkage disequilibrium was considered as short range when estimated between pairs of loci closer than 10 kilo base pairs (kbp) from each other (loci closer than 1 % on a chromosome, or ~1 centiMorgan [cM] as recombination rate was set to be constant along chromosomes). Linkage disequilibrium between loci on different chromosomes was used as a proxy for all long range LD. The decay of LD with physical distance was estimated following Hill and Weir (1988); a non-linear model was fitted between LD and distance in kbp and LD half decay distance was estimated as the distance at which LD is half of its predicted maximum value.

GWA analyses and allele frequency change following selection

The association between allelic variants of loci and phenotype were tested in the R package *GenABEL* (Aulchenko *et al.* 2007). To account for relatedness, a kinship matrix, K, was estimated by the *ibs* function, which calculates the average identity by state (IBS) for all pairs of individuals. The function *polygenic* was used to estimate residual trait variance and the inverse of the variance-covariance matrix in the presence of relatedness. Outputs from function *polygenic* were further analyzed with function *mmscore*, which implements the score test for association between genetic polymorphisms and a trait (Chen & Abecasis 2007). The *mmscore* function can be used on both quantitative and binary traits, which allowed us to (1) test allele frequency change before and after selection by treating viability as a binary response variable (binary), and

(2) directly compare this to tests for associations between genotypes and the underlying quantitative phenotypic trait under selection (quantitative). For the binary GWA analyses in the simulated data, we coded all individuals with phenotypic values larger than the mean plus 0.3 standard deviations of the mean (see above) as '1', representing individuals present in the population after selection, else they were coded as '0' (not present in the population after selection). Note that in such analyses of selection experiments one assumes that selection acts on one or more unknown but heritable trait(s), and thus that the only 'phenotypic' information needed for each individual present before selection is its presence/absence in the population also after selection.

We also performed all analyses ignoring relatedness by setting all pairwise IBS values to zero. In the absence of covariates, this reduced our GWA analyses to linear regressions. This was done for two reasons; (1) it allowed us to estimate the *p*-value inflation caused by population stratification when relatedness is not taken into account, and (2) it allowed us to compare binary GWA analyses on viability to previously used permutation tests for assessing significance of allele frequency change before and after selection. For the permutation tests, empirical null-distributions for allele frequency change before and after selection were generated by random permutation of samples as in Gompert *et al.* (2014). To avoid unnecessary replication but still achieve reasonable precision of estimated *p*-values, we continued permutations until at least 10 permuted values were more extreme than the observed, with a minimum 1000 permutations for all tests. This approach is similar to a sequential probability ratio test (Fay *et al.* 2007). Due to the large number of permutations required by the above procedure, the comparisons between binary

GWA analyses (ignoring relatedness) and the permutation tests were restricted to four data sets for each combination of levels of population structure and number of causal loci (36 data sets in total).

The amount of residual *p*-value inflation due to population stratification was estimated by regression in a Q-Q plot based on observed versus expected χ^2-values under the null-distribution. The inflation factor, λ, is the slope of the regression, where λ-values larger than one indicate *p*-value inflation. Although no strict guidelines exist, here we considered $\lambda > 1.1$ to indicate strong *p*-value inflation.

In our simulated data, we tested the correlation of *p*-values ($-\log_{10}$ transformed) from the binary and quantitative GWA analyses. A strong correlation indicates that both tests identify the same genomic regions being associated with the trait under selection. To test the extent to which the underlying population structure (rather than true genetic correlations) in the data affects the outcomes of these tests more generally, we also tested the correlation of *p*-values from binary and quantitative GWA analyses when traits were based on two different sets of causal loci. To assure independence, we did not allow any of the causal loci from the two sets to be closer than 100 kbp from each other (~10 cM). If population stratification has no influence on *p*-values, the expected number of significant correlations from such tests should be close to 5% and display no inflation (in a Q-Q plot comparing $-\log_{10}$ *p*-values) compared to *p*-values following a uniform null-distribution. To investigate how associations between genotypes and phenotypes depends on population stratification, we tested to what extent the *t*-statistics from the *p*-value correlations between binary and quantitative GWA analyses in turn correlated with the mean of $\log_{10} \lambda$

for each pair of tests. This value, $\log_{10} \bar{\lambda}$, was used as a proxy for how much population stratification there was in the data, which could vary considerably also within the different levels of population structure (note that population stratification can also be present in data sets with no population structure, see also Discussion). The same 900 simulated data sets as above were used except we dropped the number of causal loci as a factor and used four causal loci for all analyses (i.e. n = 300 for each level of population structure).

To control for multiple testing, we estimated q-values (expected proportion of false positives among all tests that are deemed significant) using the function *qvalue* from Bioconductor's *qvalue* package (Dabney & Storey 2014). We considered a test significant when q < 0.1, i.e. accepting a 10% probability that that the test is a false positive. Here we define power of a test as the average number of significant causal loci, and all significant loci further than 50 kbp (~5 cM) away from any causal loci were considered false positives. This distance is likely to be appropriate considering the distance at which LD breaks down in the simulated data set (see Results).

Artificial selection in house sparrows and SNP genotyping

An artificial selection experiment on tarsus length in two populations of house sparrows was conducted during the years 2002 to 2005 as described in Kvalnes et al. (in review). In short, for four successive years (2002-2005) ~90% of all individuals on each of two islands (Leka and Vega) in northern Norway were captured each February, during approximately two weeks. At the end of this two-week period, all individuals with a tarsus longer than the mean plus 0.3 standard deviations of the mean were released back to Leka, and individuals with a tarsus

306 shorter than the mean minus 0.3 standard deviations of the mean were released
307 back to Vega. These individuals comprised the selected individuals. The
308 remaining individuals (non-selected) were relocated to distant mainland
309 populations > 95 km's away. Thus, the strength of selection was the same as for
310 the simulated data above. Individuals were genotyped at fourteen microsatellite
311 loci to establish high quality genetic pedigrees (Rønning *et al.* 2016). Individuals
312 with the most informative family links (File S1, Supporting Information) were
313 chosen for genotyping on a custom house sparrow 10 K Illumina iSelect HD
314 BeadChip (Hagen *et al.* 2013). Of the initial 10000 SNPs, 6492 were variable, of
315 high quality and could be mapped to a reference genome (Hagen et al., in
316 preparation). This data was further filtered such that no more than 20% of
317 genotypes were missing for any locus (median < 0.1%) or individual (median =
318 0%). Loci that at some point (within an island) became fixed during the
319 experimental period were ignored, as a null-distribution for such loci for those
320 years cannot be generated. These procedures resulted in 5131 (from 267
321 individuals) and 5075 SNPs (from 273 individuals) available for analysis on the
322 island of Leka and on Vega, respectively. More detailed sample information is
323 available in File S1 (Supporting Information).
324
325 *GWA analyses and allele frequency change in house sparrows*
326 GWA analyses on tarsus length were conducted on the two islands separately
327 using the same data sets as used for testing allele frequency change. Because
328 tarsus length does not change with age, we used mean values adjusted for
329 fieldworker (Kvalnes et al. in review) when multiple measurements for adult

individuals were available (Jensen *et al.* 2003; 2008). For the function *polygenic* sex was included as fixed factor.

Allele frequency change was estimated within each island as the population mean allele frequency in all adult individuals immediately before artificial selection a given year (baseline), minus the population mean allele frequency in adult individuals present in the population directly after artificial selection (i.e. excluding the individuals that were removed from the island that year; see above). The total allele frequency change due to artificial selection for the experiment was attained by the sum of all the within-year changes. Thus, loci with large allele frequency changes in the same direction each year have the highest total allele frequency changes. Note that this only measures allele frequency change directly due to artificial selection and does not take into account the fact that drift and/or natural selection also may cause allele frequencies in the population to change between two successive artificial selection episodes. This was done to isolate the effect of the artificial selection on allele frequency change. *P*-values for allele frequency change for each locus were attained from an empirical null-distribution acquired from gene-dropping simulations (Gratten *et al.* 2012; File S2, Supporting Information). *P*-value inflation in gene-dropping simulations is likely to stem from the presence of relatedness among the founders; in the simulations founders are assumed only to be related by chance (File S2, Supporting Information). To correct for *p*-value inflation in the gene-dropping simulations, we performed GC by adjusting for λ, which was estimated directly from $-\log_{10} p$ (Price *et al.* 2010). Function *qvalue* was used to estimate *q*-values and the proportions of genes for which the null hypothesis is true ($1-\pi 0$).

Results

P-value inflation in allele frequency change before and after selection

When relatedness was ignored in the binary GWA analyses, the correlations between $-\log_{10} p$ from random permutation of samples and binary GWA analyses (both testing for allele frequency change before and after selection), were close to unity for all 36 simulated data sets (all $r_p > 0.99$). There was no significant effect of population structure ($P = 0.65$, $F_{(2,27)} = 0.44$) or number of causal loci ($P = 0.86$, $F_{(2,27)} = 0.15$) on these correlations. Thus, when ignoring relatedness, binary GWA analyses can be considered as a proxy for previously used permutation tests for assessing significance of allele frequency change before and after selection.

For both GWA testing for allele frequency differences before and after selection with viability treated as binary response variable and GWA analyses performed on the underlying quantitative trait under selection, heritability is the main prerequisite for *p*-value inflation to occur (Fig. S1, Supporting Information). Thus, we present result on heritable traits only. When ignoring relatedness, considerable *p*-value inflation existed in data sets simulated under random mating (Fig. 1 A) for both binary and quantitative GWA analyses. This *p*-value inflation increased drastically with increasing population structure (Fig. 1 B). However, accounting for relatedness greatly reduced *p*-value inflation in all cases (Fig. 1 B).

False positive rates and power to detect causal loci for binary GWA testing for allele frequency change before and after selection reflect the results of *p*-value inflation presented above and agree well with what is known for GWA

studies in general (Table 1). The main findings are as follows. In the presence of strong population structure and when relatedness was not accounted for, all tests displayed large numbers of false positives. When populations were simulated under random mating, the mean number of false positives was still large and exceeded the mean number of significant causal loci. In contrast, false positives were close to zero in all tests when accounting for relatedness and using GC to correct for any residual *p*-value inflation. The power to detect causal loci was always lower for binary GWA analyses compared to quantitative GWA analyses. Power to detect causal loci when accounting for relatedness as well as performing GC was generally low and decreased with increasing number of causal loci. For instance, with eight causal loci significant causal loci (one ore more) could only be detected in 17 out of 300 data sets (pooled over all levels of population structure).

P-value inflation was closely associated with long range LD caused by population stratification. In our simulated data sets, both the median and median absolute deviation for LD increased with population structure, at both short and long range (Fig. 2). A marked difference between short and long range LD was seen in the 95 % quantiles, where LD increased more with increasing population structure at long range (Fig. 2). Furthermore, LD half decay distance increased with increasing population structure (1.68 cM, 1.87 cM and 2.57 cM for N_em = $N_e/2$, $N_em = 2$ and $N_em = 1$, respectively). Linkage disequilibrium plotted against physical distance for all levels of population structure are shown in Fig. S2 (Supporting Information).

Do binary and quantitative GWA associate the same genomic regions with traits under selection?

There was a strong correlation between $-\log_{10} p$ from binary and quantitative GWA analyses across all data sets when tests were conducted on the same phenotypic trait (Fig. 3 A and C). These correlations were stronger when ignoring relatedness (Fig. 3 A) compared to when relatedness was accounted for (Fig. 3 C). The correlations generally increased with increasing $\log_{10} \overline{\lambda}$ (Fig. 3 A and C). When ignoring relatedness, the increase in correlation depended on population structure (Fig. 3 A) but was independent of population structure when accounting for relatedness (Fig. 3 C). This demonstrates that the underlying population stratification causes similar and strong biases in test statistics from GWA analyses testing for allele frequency change before and after selection and quantitative GWA analyses directly testing for associations between genotypes and traits under selection.

When the phenotypic traits under selection were based on different independent sets of causal loci and relatedness was ignored, 75% (inflated by a factor of 13.5 compared to a uniform null-distribution) of all correlations between $-\log_{10} p$ from quantitative and binary GWA analyses were significant (Fig. 3 B). This dropped to 56% (inflated by a factor of 7.30 compared to a uniform null-distribution) when relatedness was accounted for (Fig. 3 D). When ignoring relatedness, this correlation increased with $\log_{10} \overline{\lambda}$ for data sets with moderate and strong population structure but not for data sets simulated under random mating (Fig. 3 B). However, when relatedness was accounted for, correlations no longer increased with $\log_{10} \overline{\lambda}$ for any level of population structure. Thus, even when variation in phenotypic traits was explained by

independent sets of loci in the binary and quantitative GWA analyses, the underlying population stratification caused *p*-values from these two tests to be similarly biased.

Allele frequency change in artificially selected house sparrow populations

When testing for allele frequency change using gene-dropping simulations without GC, we found *p*-value inflation for both house sparrow populations (Fig. 4; Leka: $\lambda = 1.4$, SE=4.6×10^{4}; Vega: $\lambda = 1.1$, SE=4.9×10^{4}). Without GC, The proportions of rejected null-hypotheses were estimated to 23 % at Leka and 9.4 % at Vega. Furthermore, 33 loci were significant at $q < 0.1$ in the Leka population, while no loci were significant (i.e. had $q < 0.1$) in the Vega population. With GC, q-values for the most significant loci increased from 0.053 to 0.51 at Leka and from 0.19 to 0.49 at Vega, and proportions of rejected null-hypotheses dropped to zero in both populations. Hence, after GC no loci showed larger allele frequency change than could be expected by random genetic drift alone.

When ignoring relatedness, *p*-value inflation with quantitative GWAS on tarsus length, was high in both populations (Leka: $\lambda = 1.9$, SE = 1.5×10^{3}; Vega: $\lambda = 1.7$, SE = 1.4×10^{4}). After accounting for relatedness, λ's were below one for both populations and the q-values for the most significant loci were 0.91 and 0.97 at Leka and Vega, respectively. Hence, after accounting for relatedness, no loci were significantly associated with tarsus length.

After accounting for relatedness, $-\log_{10} p$ from GWA analyses for tarsus length and within year allele frequency change summed over the whole selection experiment (as tested by gene-dropping simulations) were significantly

correlated (Leka: $r_p = 0.29$, $t = 22$, $df = 5029$ $p < 0.001$; Vega: $r_p = 0.36$, $t = 28$, $df = 5173$, $p < 0.001$), with even stronger correlations when ignoring relatedness (Leka: $r_p = 0.52$, $t = 43$, $df = 5129$, $p < 0.001$; Vega: $r_p = 0.43$, $t = 35$, $df = 5173$, $p < 0.001$). This suggests that artificial selection on tarsus length has influenced within year allele frequency changes within both islands (but see Discussion).

Discussion

Test statistics for allele frequency change in experimental evolution and GWA studies both ultimately rely on associations between genotypes and phenotypes (Fig. 3). As such, we here show that test statistics for allele frequency change and standard GWA analyses are equally prone to *p*-value inflation (Fig. 1, 3 and 4 and Table 1). However, we also show that methods to assess the magnitude of *p*-value inflation and account for relatedness in GWA studies are also applicable for testing for significant allele frequency change in experimental evolution studies (Fig. 1 and Table 1). Two additional benefits of using previously developed GWA approaches to asses the significance of allele frequency change are reduced computational time (at least relative to previously used permutation tests) and the possibility to account for additional covariates, but this is not considered in the present paper.

In permutation tests probability estimates are subject to error due to sampling the population of possible permutations (Ojala & Garriga 2010), generating a trade-off between precision of the *p*-values and computational resources. Previous studies assessing the significance of allele frequency change before and after selection by permutation have relied on only 1000 replicates (Gompert & Buerkle 2011; Pespeni *et al.* 2013). The minimum *p*-values one can

attain from such tests is the inverse of the number of replicates (one-tailed tests), which has the potential to lead to misleading results when correcting for multiple testing (Phipson & Smyth 2010) and does not allow for proper estimation of *p*-value inflation. In contrast, current GWA methods are optimized for large data sets and in the present paper we have demonstrated that they can be used to assess the significance of allele frequency change by fitting a binary response variable e.g. present/absent after an episode of selection. This enables accurate *p*-values for association statistics to be estimated much faster.

In our empirical data set from artificial selection on tarsus length in house sparrows, we report substantial *p*-value inflation for within year allele frequency change (*p*-values were attained from null-distributions generated by gene-dropping simulations rather than binary GWA analyses). By ignoring this *p*-value inflation, a substantial proportion of our loci (23% at Leka and 9.4% at Vega) would have erroneously been thought to be (directly or indirectly) associated with causal variants underlying variation in tarsus length. While we could not directly account for relatedness when estimating the *p*-values we could still perform GC. In doing so the expected number of significant loci dropped to zero in both populations. Hence, we emphasize that when testing for significance of allele frequency change, even in complex experimental designs spanning multiple generations, *p*-value inflation is an important confounding factor that potentially can be addressed with GC.

Power to detect loci under selection in experimental evolution studies

The power to detect causal loci in GWA studies is largely determined by the number of causal loci, the difference in phenotypic values between alternative

allelic variants, and the degree of heterozygosity (Martin & Jiggins 2001; Korte & Farlow 2013). From a statistical perspective, quantitative traits are preferred over binary (case/control) because they improve power to detect a genetic effect (Bush & Moore 2012). This is also reflected here where the power of binary GWA analyses testing for allele frequency change before and after selection was always lower than quantitative GWA analyses performed directly on the underlying phenotypic trait under selection (Table 1).

Our simulated data were designed to mimic artificial selection experiments, where the selected phenotype is known and precise cut-off values for truncated selection can be used. The only variation with respect to survival of a particular phenotype in our simulations was environmental, specifically determined by the heritability of the trait under selection. In contrast, in natural selection (experiments) the researcher has no control over individual survival. As natural selection is subject to stochasticity, this generates additional variation (on top of environmental) with respect to the survival of a particular phenotype. Thus, we predict that the power to detect causal loci from test statistics for allele frequency change under natural selection (experiments) to be even lower than shown here.

Linkage disequilibrium

False statistical associations between genotypes and phenotypes are ultimately caused by long range LD in both GWA studies (Korte & Farlow 2013) and experimental evolution (Tobler *et al.* 2014; Schlötterer *et al.* 2015). Many biological processes, in particular mating among relatives (at any level of the population hierarchy) initially increase LD between loci across the whole

genome (Charlesworth & Charlesworth 2010; Kemppainen *et al.* 2015). Nevertheless, independent segregation and assortment of chromosomes ensures along with recombination that LD typically extends only short physical distances within chromosomes in large natural populations at any given time (Charlesworth & Charlesworth 2010). However, the fact that decay of LD can only take place in the presence of recombination that requires mating between individuals is often overseen. Thus, when the study sample comprises individuals from different populations (that do not meet to potentially mate), admixture LD, that is completely independent of physical distance, is created that will not decay with time (Fig. 2 and Fig. S2, Supporting information; Charlesworth & Charlesworth 2010; Kemppainen *et al.* 2015). This is the type of LD that is present in our simulated data with moderate and strong population structure. However, even in panmictic populations LD can be strong between physically distant pairs of loci due to genetic drift, selection and other sampling effects (particularly if N_e is small or only a few individuals have been selected or sampled; Charlesworth & Charlesworth 2010). This is evident from our data sets simulated under random mating despite large effective population sizes (N_e=10000). When ignoring relatedness, long range LD was sufficient to cause at least one false positive in 37% of the data sets (Table 1), and 82% of all tests showed strong *p*-value inflation in the binary GWA analyses testing for allele frequency change before and after selection (see also Fig. 1). This was most likely because even in such cases there is variation in relatedness between individuals (i.e. all individuals are not equally related, or unrelated, to each other), which cause some population stratification in the data that is not easily detected by common population genetic tools. In other words, even in studies where

individuals are randomly sampled from large and arguably panmictic populations, *p*-value inflation in test statistics for allele frequency change may still be present (see also Tobler *et al.* 2014 and Schlötterer *et al.* 2015).

Population stratification has strong influence on test statistics for allele frequency change in experimental evolution studies

It has been suggested that candidate genes from experimental evolution can be validated by GWA studies (Tobler *et al.* 2014; Schlötterer *et al.* 2015). In our simulated data *p*-values from quantitative and binary GWA analyses were much more correlated than expected by chance, when tests were conducted on the same data set but when the phenotypes were based on different and independent sets of causal loci (Fig. 3 B). Thus, here the correlations were caused by the underlying LD structure due to population stratification in the data rather than due to real genetic correlations, and this also occurred in randomly mating populations. Accounting for relatedness in both the quantitative and binary GWA analyses alleviated this to some extent (Fig. 3 D). Nevertheless, in data sets simulated under random mating, *p*-values were still inflated by a factor of 7.3 (compared to a null-distribution of no effect) resulting in significant *p*-value correlations in 56% of the data sets (Fig. 3 D).

It has been argued that due to allele frequency variation and possible epistatic interactions "lack of replication does not necessarily indicate lack of an effect", if these tests are performed on different data sets (Schlötterer *et al.* 2015). It is clear that the null-distribution of no effect when comparing *p*-values from allele frequency change and GWA analyses (on the trait under selection) does not lead to a uniform distribution of *p*-values. Instead it depends on the

579 genetic architecture of the data and the underlying population stratification.
580 These were known for our simulated data and thus the results in Fig. 3 (C and D)
581 can be considered as empirical null-distributions for the results in Fig. 3 (A and
582 B). When a null-distribution cannot be created, the safest way to remove
583 confounding effects of population stratification when validating candidate loci
584 under selection with GWA studies is indeed to perform these tests on data sets
585 from two different populations.

586 In experimental evolution, it is usually argued that parallel allele
587 frequency changes in replicated selection experiments are the signature of
588 selection (Tobler *et al.* 2014; Gompert *et al.* 2014; Schlötterer *et al.* 2015).
589 However, following the argumentation above, if individuals in replicated
590 selection experiments are sampled from the same population, the same
591 underlying population stratification is likely to be present also among the
592 individuals in the replicated experiments. This, in turn, may cause correlated
593 allele frequency changes due to relatedness (long range LD) rather than true
594 associations between the loci and causal genetic variants affecting the trait.
595 However, also here the different methods to assess and correct for *p*-value
596 inflation developed for GWA studies can potentially be used. In addition, to
597 increase independence between replicated experiments, individuals could be
598 collected from different populations, with the caveat that different causal
599 variants then may be responsible for the traits in these populations.

600 In our artificial selection experiment $-\log_{10} p$ from GWA analyses on
601 tarsus length and allele frequency change were strongly correlated, even after
602 accounting for relatedness. Thus, the allele frequency changes we observed were
603 most likely due to the artificial selection on tarsus length that we imposed on the

604 populations. However, due to population stratification and the lack of a proper
605 null-distribution (as argued above) we cannot exclude completely the possibility
606 that the correlations we observed were caused by the underlying population
607 stratification rather than selection on casual genetic variants affecting tarsus
608 length. However, the fact that considerable *p*-value inflation in test statistics for
609 allele frequency change existed (in particular in Leka; Fig. 4) suggests that
610 evolutionary change in a heritable trait (or traits) indeed had occurred (see
611 Kvalnes et al., in review). Nevertheless, we could not determine if any of the loci
612 were associated with these traits, except through long range LD caused by
613 population stratification in the data.
614
615 *Biological consequences of selection in the presence of population stratification*
616 In GWA studies *p*-value inflation is predominantly a statistical issue, i.e. it may
617 lead to false claims of association between loci and the trait of interest. However,
618 it should be recognized that allele frequency change due to selection in stratified
619 populations (that causes *p*-value inflation) could have biological implications as
620 well. If individuals with higher survival rates or reproductive success are more
621 closely related than expected by chance (i.e. fitness depends on a heritable trait),
622 any alleles that are identical by decent among the selected individuals are likely
623 to hitchhike to higher frequency along with any causal variants for that trait. In
624 natural populations, the biological consequences of this can for instance be; (1)
625 reduced N_e (regardless of any eventual change in the census population size [N_c])
626 and as a consequence increased drift and rates of population differentiation, (2)
627 inbreeding, (3) maladaptation and (4) reduced evolutionary potential. Below we

provide biological examples for scenarios 1-3. Evidence for scenario (4) follows indirectly from point (1).

(1) Exceptionally fast population differentiation was detected between geographically proximate populations of trout (*Salmo trutta*) that had undergone rapid adaptation to heavy metal contamination, relative to pristine populations much further apart (Paris *et al.* 2015). A reduction in N_c, with a corresponding reduction in N_e, could alone explain the fast drift within these populations. However, due to very strong selection, it is likely that the amount of drift was stronger than what could have been predicted solely by the reduction in N_c. According to our findings, this is particularly likely if selection operated on a highly heritable trait (possibly controlled by few genes of large effect) and populations exhibited strong population stratification before selection.

(2) Strong selection on heritable traits can directly lead to inbreeding, as then by definition individuals in the subset of the population that survives are likely to be more related to each other than expected by chance. Support for this comes from a study on an insular population of song sparrows (*Melospiza melodia;* Keller *et al.* 2001). This study showed that survival following a severe storm was not only higher for individuals with long wings (selection) but also for individuals with high inbreeding coefficients. Indirect evidence for this comes also from our artificial selection experiments in house sparrows presented here. The very existence of *p*-value inflation for allele frequency change suggests that non-random sampling with respect to relatedness between individuals indeed had occurred. Inbreeding potentially leads to inbreeding depression via increased genetic load (Charlesworth & Willis 2009), so strong selection on

heritable traits may have severe immediate negative consequences for the survival of the affected populations.

(3) Selection on heritable traits can lead to maladaptation when sub-optimal genotypes hitchhike to higher frequency due to LD (including long range LD e.g. caused by relatedness among the selected individuals) with loci under selection. That artificial selection in small populations can lead to maladaptation is already well known in commercial breeding programs (Garland 2003). This can e.g. clearly be seen in dogs where selective breeding has led to accumulation of negative mutations causing high prevalence of diseases in certain breeds (Marsden *et al.* 2016).

In other words, the potential biological consequences of strong selection in natural populations may have more important implications for conservation management strategies than previously recognized. This is expected to be especially true for, but not limited to, populations with strong population stratification.

Conclusions

As proof of concept, we have shown with simulated data that test statistics for allele frequency change before and after selection behave similarly to those from GWA studies on quantitative traits. Thus, the approaches and methods already available for GWA studies to account for relatedness and correct for *p*-value inflation is available also to experimental evolution studies. We emphasize that for any test statistic that ultimately depends on associations between genotypes and phenotypes, the potential of *p*-value inflation has to be considered and properly dealt with. Here we provide two examples of how this can be done:

677 binary GWA analyses when including relatedness as a random effect, and
678 genomic control. Importantly, our study also shows using both simulations and
679 empirical data from an artificial selection experiment in two free-living bird
680 populations, that allele frequencies in large parts of the genome may change
681 when selection is acting on a heritable trait. These genetic changes are likely to
682 have considerable and wider consequences for the eco-evolutionary dynamics of
683 such populations in the immediate future.
684
685 **Acknowledgements**
686 We are grateful to the inhabitants and farmers on Leka and Vega, whose
687 hospitality enabled us to collect the empirical data used in this study. We thank
688 R. Dahl, O.R. Davidsen, F. Jørgensen, T. Kolaas, L.K. Larsen, A. Lorås, P.A.
689 Martinussen, R. Moe, M. Mørkved, R. Rismark, B.G. Stokke, and K. Sørensen for
690 help during the field work, and Å.A. Borg Pedersen, O.R. Davidsen, R. Røsbak, and
691 K. Yttersian Sletta for help with the laboratory work. We also thank Rowan
692 Barrett who originally suggested using binary GWA analyses to assess the
693 significance of allele frequency change. The empirical research was carried out in
694 accordance with permits from the Norwegian Animal Research Authority
695 (permits S-2007/1482 and ID-4011) and the Ringing Centre at Stavanger
696 Museum, Norway. The study was supported by grants from the Research Council
697 of Norway, project number 221956, and Strategic University Program (SUP) in
698 Conservation Biology, project number 204303. This work was also partly
699 supported by the Research Council of Norway through its Centres of Excellence
700 funding scheme, project number 223257.
701

Data Accessibility

Data and R-code available from the Dryad Digital Repository:

http://dx.doi.org/10.5061/dryad.vg4fj

Author contributions

PK, HJ, BES, THR, BR, IJH and TK designed the project. PK executed all analyses and simulations. THR and HJ did most of the fieldwork for the artificial selection experiment. IJH, AMB, SL, HJ and AH generated SNP and reference genome data for the artificial selection experiment. PK, BR and HJ wrote the paper and all authors contributed with comments on the manuscript.

References

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics (Oxford, England)*, **23**, 1294–1296.

Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, **12**, 767–780.

Brakefield PM (2003) Artificial selection and the development of ecologically relevant phenotypes. *Ecology*, **84**, 1661–1671.

Burke MK, Dunham JP, Shahrestani P *et al.* (2010) Genome-wide analysis of a long-term evolution experiment with Drosophila. *Nature*, **467**, 587–590.

Bush WS, Moore JH (2012) Chapter 11: Genome-wide association studies. *PLoS computational biology*, **8**, e1002822.

Charlesworth B, Charlesworth D (2010) *Elements of evolutionary genetics | Clc*. Roberts and Company Publishers, Greenwood Village, Colorado.

Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. *Nature Reviews Genetics*, **10**, 783–796.

Chen W-M, Abecasis GR (2007) Family-based association tests for genomewide association scans. *American journal of human genetics*, **81**, 913–926.

Clayton DG, Walker NM, Smyth DJ *et al.* (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature genetics*, **37**, 1243–1246.

Conner JK (2003) Artificial selection: a powerful tool for ecologists. *Ecology*, **84**, 1650–1660.

Dabney A, Storey JD (2014) qvalue: Q-value estimation for false discovery rate control. *R package version 1.43.0*.

Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

740 Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to
741 genetic-based association studies. *Theoretical population biology*, **60**, 155–
742 166.
743 Dudley JW, Lambert RJ, La Roche De IA (1977) Genetic Analysis of Crosses
744 among Corn Strains Divergently Selected for Percent Oil and Protein1. *Crop*
745 *Science*, **17**, 111.
746 Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in natural
747 populations. *Nature*, **452**, 169–175.
748 Excoffier L, Foll M (2011) fastsimcoal: a continuous-time coalescent simulator of
749 genomic diversity under arbitrarily complex evolutionary scenarios.
750 *Bioinformatics (Oxford, England)*, **27**, 1332–1334.
751 Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust
752 demographic inference from genomic and SNP data. *PLoS Genetics*, **9**,
753 e1003905.
754 Falconer DS, Mackay TFC (1996) *Introduction to Quantitative Genetics*. Longman,
755 Essex.
756 Fay MP, Kim H-J, Hachey M (2007) On Using Truncated Sequential Probability
757 Ratio Test Boundaries for Monte Carlo Implementation of Hypothesis Tests.
758 *Journal of Computational and Graphical Statistics*, **16**, 946–967.
759 Garland T (2003) *Selection experiments: an under-utilized tool in biomechanics*
760 *and organismal biology* (IL Bels, J-P Gasc, A Casinos, Eds,). BIOS Scientific
761 Publishers Ltd, Oxford, Oxford.
762 Gompert Z, Buerkle CA (2011) A hierarchical Bayesian model for next-generation
763 population genomics. *Genetics*, **187**, 903–917.
764 Gompert Z, Comeault AA, Farkas TE *et al.* (2014) Experimental evidence for
765 ecological selection on genome variation in the wild. *Ecology letters*, **17**, 369–
766 379.
767 Gratten J, Pilkington JG, Brown EA *et al.* (2012) Selection and microevolution of
768 coat pattern are cryptic in a wild population of sheep. *Molecular ecology*, **21**,
769 2977–2990.
770 Gromko MH (1995) Unpredictability of Correlated Response to Selection:
771 Pleiotropy and Sampling Interact. *Evolution*, **49**, 685.
772 Gromko MH, Briot A, Jensen SC, Fukui HH (1991) Selection on Copulation
773 Duration in Drosophila melanogaster: Predictability of Direct Response
774 Versus Unpredictability of Correlated Response. *Evolution*, **45**, 69.
775 Hagen IJ, Billing AM, Rønning B *et al.* (2013) The easy road to genome-wide
776 medium density SNP screening in a non-model species: development and
777 application of a 10 K SNP-chip for the house sparrow (Passer domesticus).
778 *Molecular ecology resources*, **13**, 429–439.
779 Hao K, Di X, Cawley S (2007) LdCompare: rapid computation of single- and
780 multiple-marker r2 and genetic coverage. *Bioinformatics (Oxford, England)*,
781 **23**, 252–254.
782 Heidaritabar M, Vereijken A, Muir WM *et al.* (2014) Systematic differences in the
783 response of genetic variation to pedigree and genome-based selection
784 methods. *Heredity*, **113**, 503–513.
785 Hill WG, Caballero A (1992) Artificial selection experiments. *Annual Review of*
786 *Ecology and Systematics*, **23**, 287–310.
787 Hill WG, Weir BS (1988) Variances and covariances of squared linkage
788 disequilibria in finite populations. *Theoretical population biology*, **33**, 54–78.

789 Illingworth CJR, Parts L, Schiffels S, Liti G, Mustonen V (2012) Quantifying
790 selection acting on a complex trait using allele frequency time series data.
791 *Molecular Biology and Evolution*, **29**, 1187–1197.
792 Jensen H, Saether BE, Ringsby TH *et al.* (2003) Sexual variation in heritability
793 and genetic correlations of morphological traits in house sparrow (Passer
794 domesticus). *Journal of evolutionary biology*, **16**, 1296–1307.
795 Jensen H, Steinsland I, Ringsby TH, Saether B-E (2008) Evolutionary dynamics of
796 a sexual ornament in the house sparrow (Passer domesticus): the role of
797 indirect selection within and between sexes. *Evolution*, **62**, 1275–1293.
798 Keller LF, Jeffery KJ, Arcese P *et al.* (2001) Immigration and the ephemerality of a
799 natural population bottleneck: evidence from molecular markers.
800 *Proceedings. Biological sciences / The Royal Society*, **268**, 1387–1394.
801 Kemppainen P, Knight CG, Sarma DK *et al.* (2015) Linkage disequilibrium
802 network analysis (LDna) gives a global view of chromosomal inversions,
803 local adaptation and geographic structure. *Molecular ecology resources*, **15**,
804 1031–1045.
805 Korte A, Farlow A (2013) The advantages and limitations of trait analysis with
806 GWAS: a review. *Plant methods*, **9**, 29.
807 Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human
808 population structure on large genetic association studies. *Nature genetics*,
809 **36**, 512–517.
810 Marsden CD, Ortega-Del Vecchyo D, O'Brien DP *et al.* (2016) Bottlenecks and
811 selective sweeps during domestication have increased deleterious genetic
812 variation in dogs. *Proceedings of the National Academy of Sciences*, **113**, 152–
813 157.
814 Martin SH, Jiggins CD (2001) *Genomic Studies of Adaptation in Natural*
815 *Populations*. John Wiley & Sons, Ltd, Chichester, UK.
816 McCarthy MI, Abecasis GR, Cardon LR *et al.* (2008) Genome-wide association
817 studies for complex traits: consensus, uncertainty and challenges. *Nature*
818 *Reviews Genetics*, **9**, 356–369.
819 Nielsen R (2005) Molecular Signatures of Natural Selection. *Annu Rev Genet*, **39**,
820 197–218.
821 Ojala M, Garriga GC (2010) Permutation Tests for Studying Classifier
822 Performance. *The Journal of Machine Learning Research*, **11**, 1833–1863.
823 Palmer JO, Dingle H (1986) Direct and Correlated Responses to Selection Among
824 Life-History Traits in Milkweed Bugs (Oncopeltus fasciatus). *Evolution*, **40**,
825 767.
826 Pardo-Diaz C, Salazar C, Jiggins CD (2015) Towards the identification of the loci
827 of adaptive evolution. *Methods in ecology and evolution / British Ecological*
828 *Society*, **6**, 445–464.
829 Paris JR, King RA, Stevens JR (2015) Human mining activity across the ages
830 determines the genetic structure of modern brown trout (Salmo trutta L.)
831 populations. *Evolutionary applications*, **8**, 573–585.Parts L, Cubillos FA,
832 Warringer J *et al.* (2011) Revealing the genetic structure of a trait by
833 sequencing a population under selection. *Genome research*, **21**, 1131–1138.
834 Pespeni MH, Sanford E, Gaylord B *et al.* (2013) Evolutionary change during
835 experimental ocean acidification. *Proceedings of the National Academy of*
836 *Sciences*, **110**, 6937–6942.
837 Phipson B, Smyth GK (2010) Permutation P-values should never be zero:

calculating exact P-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, **9**, Article39.

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, **11**, 459–463.

Remolina SC, Chang PL, Leips J, Nuzhdin SV, Hughes KA (2012) Genomic basis of aging and life-history evolution in Drosophila melanogaster. *Evolution*, **66**, 3390–3403.

Rønning B, Broggi J, Bech C, Moe B (2016) Is basal metabolic rate associated with recruit production and survival in free-living house sparrows? *Functional Ecology*, 1140–1148.

Schlötterer C, Kofler R, Versace E, Tobler R, Franssen SU (2015) Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. *Heredity*, **114**, 431–440.

Tobler R, Franssen SU, Kofler R *et al.* (2014) Massive habitat-specific genomic response in D. melanogaster populations during experimental evolution in hot and cold environments. *Molecular Biology and Evolution*, **31**, 364–375.

Turner TL, Miller PM (2012) Investigating natural variation in Drosophila courtship song by the evolve and resequence approach. *Genetics*, **191**, 633–642.

Turner TL, Miller PM, Cochrane VA (2013) Combining genome-wide methods to investigate the genetic complexity of courtship song variation in Drosophila melanogaster. *Molecular Biology and Evolution*, **30**, 2113–2120.

Turner TL, Stewart AD, Fields AT, Rice WR, Tarone AM (2011) Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in Drosophila melanogaster. *PLoS Genetics*, **7**, e1001336.

Zhou D, Udpa N, Gersten M *et al.* (2011) Experimental selection of hypoxia-tolerant Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, **108**, 2349–2354.

Table 1. Number of false positives and power of binary and quantitative GWA analyses for simulated data with two, four or eight causal loci. "K" indicates weather relatedness was included as a random effect. "GC" indicates if genomic control was performed. "Population structure" indicates if data was simulated with (Yes: $N_e m = 1$, i.e. 'strong population structure') or without (No: data simulated under random mating) population structure. "False Positives" is the mean (SD) number of significant loci further than 50 kbp (~5 cM) away from the closest causal locus and "Power" is the mean (SD) number of significant causal loci. * Indicates that the test is equivalent to permutation test for allele frequency change before and after selection, within a single generation. See text for details on the simulations.

K	GC	Population Structure	Number of Causal Loci	Binary GWA analyses		Quantitative GWA analyses	
				False Positives	Power	False Positives	Power
No	No	No	2	5.08 (11.78)*	0.98 (0.71)*	18.59 (27.77)	1.73 (0.60)
No	No	Yes	2	893.12 (1349.58)*	1.35 (0.77)*	1303.49 (1462.13)	1.82 (0.46)
No	Yes	No	2	0.03 (0.22)	0.38 (0.53)	0.09 (0.38)	1.22 (0.66)
No	Yes	Yes	2	0.00	0.21 (0.41)	0.00	0.41 (0.62)
Yes	No	No	2	0.18 (0.63)	0.61 (0.62)	0.66 (1.44)	1.55 (0.64)
Yes	No	Yes	2	0.39 (1.80)	0.57 (0.62)	0.71 (1.92)	1.34 (0.70)

884	Yes	Yes	No	2	0.05 (0.26)	0.47 (0.56)	0.20 (0.65)	1.42 (0.65)
885	Yes	Yes	Yes	2	0.07 (0.36)	0.42 (0.55)	0.12 (0.66)	1.20 (0.71)
886	No	No	No	4	4.83 (11.44)*	0.77 (0.92)*	16.78 (24.58)	2.32 (0.96)
887	No	No	Yes	4	1024.03 (1290.4)*	1.96 (1.56)*	1388.47 (1436.51)	2.81 (1.23)
888	No	Yes	No	4	0.07 (0.46)	0.15 (0.41)	0.00	0.40 (0.67)
889	No	Yes	Yes	4	0.01 (0.10)	0.04 (0.20)	0.00	0.11 (0.37)
890	Yes	No	No	4	0.46 (1.90)	0.32 (0.55)	0.18 (0.59)	1.08 (0.92)
891	Yes	No	Yes	4	0.23 (0.97)	0.22 (0.48)	0.51 (1.67)	0.83 (0.99)
892	Yes	Yes	No	4	0.10 (0.44)	0.22 (0.44)	0.05 (0.30)	0.74 (0.80)
893	Yes	Yes	Yes	4	0.11 (0.67)	0.15 (0.39)	0.18 (0.93)	0.44 (0.70)
894	No	No	No	8	4.46 (13.19)*	0.47 (0.89)*	13.81 (26.10)	1.49 (1.48)
895	No	No	Yes	8	709.24 (1054.34)*	2.46 (2.88)*	1041.55 (1179.46)	3.85 (2.83)
896	No	Yes	No	8	0.00	0.06 (0.31)	0.14 (0.98)	0.13 (0.42)
897	No	Yes	Yes	8	0.00	0.02 (0.14)	0.01 (0.10)	0.09 (0.38)
898	Yes	No	No	8	0.04 (0.24)	0.13 (0.39)	0.29 (1.51)	0.34 (0.77)

Yes	No	Yes	8	0.01 (0.10)	0.05 (0.22)	0.18 (1.20)	0.24 (0.53)
Yes	Yes	No	8	0.01 (0.10)	0.08 (0.31)	0.05 (0.41)	0.20 (0.59)
Yes	Yes	Yes	8	0.01 (0.10)	0.04 (0.20)	0.11 (1.10)	0.17 (0.47)

Figure 1. Violin plots for *p*-value inflation as estimated by λ for GWA analyses on viability and on the underlying quantitative trait. When testing for allele frequency change before and after selection, viability was treated as a binary response variable (Binary). GWA analyses were also performed on the underlying quantitative trait under selection (Quantitative). Analyses were performed for three levels of population structure (Random mating, Moderate and Strong) when the trait was heritable (h^2=0.5) and when (A) accounting for relatedness or (B) ignoring relatedness (see text for details). Binary GWA analyses not accounting for relatedness (B) is here used as proxy for previously used permutation tests for testing allele frequency change before and after selection. The dashed lines indicate λ = 1.1, above which we here consider *p*-value inflation to be strong.

Figure 2. Summary statistics of LD from simulated data with different levels of population structure. Statistics are shown for pairwise values of r^2 between 500 randomly chosen loci from each simulated data set (n = 300). Results are shown for LD at short range (< 10 kbp, ~1 cM) and between loci on different chromosomes as a proxy for all long range LD.

Figure 3. Correlation between $-\log_{10} p$ from quantitative and binary GWA analyses depends on $\log_{10} \bar{\lambda}$. The *t*-statistic refers to correlations between $-\log_{10}$ *P* values between quantitative and binary GWA analyses when based on the same (A, C) or different (B, D) sets of causal loci. $\text{Log}_{10} \bar{\lambda}$ is the mean inflation factor for the $-\log_{10} p$ –values from each of the tests. In the upper panel (A, B)

929 individuals' relatedness is not taken into account while in the lower panel (C, D)
930 relatedness (IBS) between all pairs of individuals was included as a random
931 effect. Correlation coefficients (r_p) are given in the figure and colored according
932 to degree of population structure. The r_p in black represents all data points
933 pooled. All significant tests (α=0.05; indicated by *) have p<0.01. The vertical
934 dashed line indicates the t-statistic for significance level p = 0.05 (with df = 4998)
935 for the original correlation test between $-\log_{10} p$ between binary and
936 quantitative GWA analyses.
937
938 Figure 4. Q-Q plot for expected versus observed $-\log_{10} p$ from within year allele
939 frequency change due to artificial selection on tarsus length. The slopes for
940 regression lines equals λ. The dashed black line indicates a line with slope = 1
941 and intercept = 0, and is shown for comparison.