Genome-Wide Mapping of Gene–Phenotype Relationships in Experimentally Evolved Populations

Laurence D. Mueller,*[1] Mark A. Phillips,[1] Thomas T. Barter,[1] Zachary S. Greenspan,[1] and Michael R. Rose[1]

[1]Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA

***Corresponding author:** E-mail: ldmuelle@uci.edu.

Associate editor: Miriam Barlow

Abstract

Model organisms subjected to sustained experimental evolution often show levels of phenotypic differentiation that dramatically exceed the phenotypic differences observed in natural populations. Genome-wide sequencing of pooled populations then offers the opportunity to make inferences about the genes that are the cause of these phenotypic differences. We tested, through computer simulations, the efficacy of a statistical learning technique called the "fused lasso additive model" (FLAM). We focused on the ability of FLAM to distinguish between genes which are differentiated and directly affect a phenotype from differentiated genes which have no effect on the phenotype. FLAM can separate these two classes of genes even with relatively small samples (10 populations, in total). The efficacy of FLAM is improved with increased number of populations, reduced environmental phenotypic variation, and increased within-treatment among-replicate variation. FLAM was applied to SNP variation measured in both twenty-population and thirty-population studies of *Drosophila* subjected to selection for age-at-reproduction, to illustrate the application of the method.

Key words: **experimental evolution, statistical learning, computer simulation.**

Introduction

In his classic work on the genetic basis of evolutionary change, Lewontin (1974) laid out the important goal of understanding the relationship between genes and phenotypes, across the entire genome. With the advent of genome-wide sequencing, evolutionary biology is now on the threshold of achieving the goal set out by Lewontin.

Most traits of interest to evolutionary biologists, from viability to physiological performance, are quantitative traits potentially affected by many genes. Quantitative trait loci ("QTL") mapping was an early attempt to identify the genetic basis of quantitative traits (Lander and Botstein 1989), but its chief strength was in identifying genomic regions of moderate to large effects that were differentiated between two inbred lines. Thus, the method was not generally applicable. Now that we can readily sequence the genomes of individuals from outbred populations, as well as sequence pooled samples of multiple individuals from such populations, we have more powerful alternative methods. There are now sophisticated linear mixed effects models for predicting phenotypes when genomes of large samples of individuals are available (Meuwissen et al. 2001; de los Campos et al. 2013; Speed and Balding 2014; Weissbrod et al. 2016).

One advantage of using model organisms subjected to sustained experimental evolution is that the levels of phenotypic differentiation can dramatically exceed the phenotypic differences observed in natural populations (Garland and Rose 2009). These strong signals should in principle make it easier to determine the genetic basis of these differences. The disadvantage is that it is currently costly to sequence many individuals from numerous populations of model organisms like fruit flies. Instead, genetic data often comes in the form of pooled sequence data ("Pool-Seq"), which allows us to estimate population-wide allele frequencies. Pool-seq coupled with experimental evolution makes the unit of observation an entire population. In many laboratories, however, there are often only three control and three selected populations available for analysis.

In addition to having a very small number of independent whole-population observations, adaptation in response to any experimental evolution regime may entail multiple phenotypes changing due to the evolution of many loci across the genome. Thus, determining the genetic foundations of a specific phenotype is going to be more complicated than simply looking for all differentiated genes arising in a particular experimental evolution paradigm. In developing techniques for analyzing pooled sequence genome-wide data, we need to sift out differentiated *but noncausal* loci when inferring the genes that affect a particular phenotype.

We test a new method called the "fused lasso additive model" (or "FLAM," Petersen et al. 2016), determining its suitability for inferring which loci are causally related to specific types of phenotypic differentiation produced by experimental evolution. We simulate multiple populations separated into groups with a highly differentiated phenotype under the control of loci with differing effects on the phenotype. The genetics for each simulated population also include loci that are highly differentiated, but do not affect the phenotype of interest, as well as a large number of undifferentiated loci.

The basic question we address here is how well the FLAM method does at identifying causal loci and sorting out noncausal loci. Although phenotype prediction is not a primary concern, we also look at the ability of the method to predict phenotypes from genomic data, so that we can compare this technique to others which also use extensive individual genomic data. Finally, we apply the method to a group of 30 *Drosophila* populations that have been adapted to environments with different ages-at-reproduction, populations for which we have both genome-wide single nucleotide polymorphism (SNP) data as well as phenotypic data.

New Approaches

Laboratory natural selection often results in large scale genetic and phenotypic differentiation. However, it is rare that only a single phenotype changes. While identifying the genetic differences between these differentiated populations is relatively straightforward, identifying which genes affect which phenotypes is more challenging.

Here, we present a statistical learning tool called the FLAM. This technique minimizes cross-validation error to find a subset of genomic SNPs which can be used to predict phenotypes. Since the penalty function that is minimized by FLAM does not produce a unique solution, we probe these solutions by repeating the minimization process from 100 different starting places and then look for those SNPs that appear most often across all FLAM results. We call these SNPs "the sparse set." To be included in the sparse set, a SNP must appear with a frequency of $k \times 100\%$ of the most frequent SNP, where k varies from 0 to 1.

Results

The results for simulated SNP databases were generated with five population-sample sizes: 6, 10, 20, 40, and 60. A single phenotype and 2,000 SNPs were measured in each population. For each of these cases, we created 100 independent databases of phenotypes and genetic data. Although FLAM will eliminate loci that don't improve the cross-validation error, it is best if some filtering can be done prior to the FLAM analysis. In all the simulations described below, we did this by comparing the allele frequencies in the replicate populations with the lowest phenotypic values to those in the replicated populations with the highest and then only including those loci that show significant differences after applying the Benjamini–Hochberg multiple testing criteria (Benjamini and Hochberg 1995). Significant differences in allele frequencies were determined by the Cochran–Mantel–Haenszel test (Landis et al. 1978).

Effects of Sample Size and Phenotypic Variation

In the simulations, databases were constructed from the linearly increasing allele frequencies (fig. 1*a* and *b*) and k was set to 0.5. The ability of FLAM to detect causative loci increases in proportion to the total number of populations (fig. 2). The simulations incorporate a "phenotypic variance" that is due to uncontrolled environmental factors and experimental techniques, in addition to the part of the variance that is due to genetic differentiation. When this phenotypic variance is low, the frequency of noncausative differentiated loci is also kept low (fig. 2*b*). With increasing phenotypic variance, the ability of FLAM to eliminate these two noncausal classes of loci is reduced (fig. 2*a*). We explore why this might be so in the next section. The neutral loci are mostly eliminated by the presorting tests and the few neutral loci that remain are then often eliminated by FLAM.

There are no substantial differences between the strong, moderate, and weak causal loci with respect to their detection (fig. 2). While this seems anomalous, it can be understood by recalling that the pattern of allele frequency change among all three groups is exactly the same. Thus, weak causative loci are picked up because their allele frequency change is in the same direction as the strong loci which are largely controlling the phenotypic changes. Later we explore the effects of causative loci that have allele frequencies that change with different patterns.

Within-Regime Genetic Differentiation

The only way that FLAM can distinguish the causal loci from the noncausal differentiated loci is if there is between-replicate genetic variation within well-differentiated groups of populations that leads to small differences in average phenotypes among the replicate populations of such groups. To illustrate this idea, we created another set of twenty-population databases which had the same level of differentiation between the high and low populations, but all populations within the high phenotype group had the same mean allele frequency, and likewise for the low phenotype group. Without between-replicate genetic variation within the major groupings, FLAM is unable to distinguish the causal from the noncausal loci (fig. 3).

Different Causal Patterns of Allele Frequency Variation

In the previous simulations, allele frequencies changed in the same fashion across subpopulations in the strong, moderate, and weak causal loci. In a second set of simulations, we let one of the causal groups have effects following the plateau patterns shown in figure 1*c* and *d*, whereas the remaining causal loci retained the patterns shown in figure 1*a* and *b*. When the strong or moderate causal loci show plateau patterns, they are still picked up in the sparse lists (fig. 4*a* and *b*), but the weak loci appear no more often than the differentiated noncausal loci (fig. 4*c*). Presumably the weak causal loci have such a minor effect on the phenotypic variation that FLAM can't detect their contribution, when their causal patterns differ from those of the strong and moderate loci. In all cases, the noncausal differentiated loci appear in the sparse lists more often when the causal loci have different patterns relative to other causal loci (fig. 4), compared with cases when they all change in the same fashion (fig. 2).

Sparse-List Selection Criteria

All the previous simulations used the criterion that loci in the sparse list must exceed kC_{max} where $k = 0.5$ and C_{max} is the count of the most frequent SNP among the 100 simulations. We next studied the effects of varying k by allowing it to range

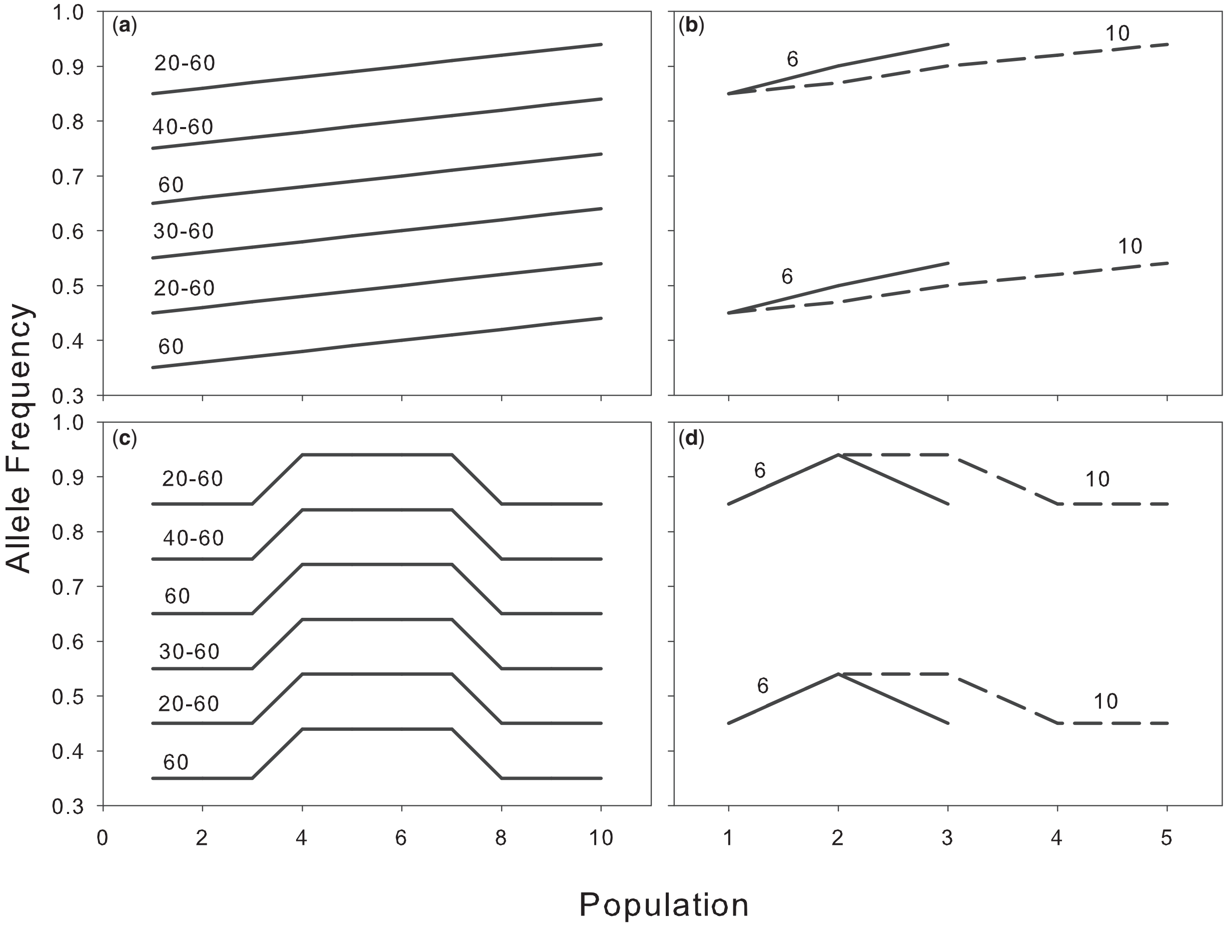


Fig. 1. Allele frequency variation used for computer simulations. Simulations used 6, 10, 20, 30, or 60 total populations. A “20–60” above the line means that these allele frequencies were used in simulations with 20, 30, and 60 populations and so on. (*a*) and (*b*) are allele frequencies used when all the causative loci show a linear increase in allele frequencies from subpopulation 1 to 10. (*c*) and (*d*) show patterns that were used in combination with the linear patterns to explore the effects of allele frequencies at causative loci changing in different patterns across the subpopulations.

over the values 0.5, 0.4, 0.3, and 0.2. We applied these criteria to the simulated results of figure 2*b*, but only show the results for the strong causal loci and differentiated noncausal loci (fig. 5). As *k* decreases, the frequency of the strong causal loci in the sparse list increases (fig. 5*a*). But the frequency of incorrectly inferring the causal involvement of differentiated noncausal loci also increases (fig. 5*b*).

When the strong causal loci had a plateau in allele frequencies, there was a pronounced increase in differentiated noncausal loci in the sparse list (fig. 4*a*). For that reason, we ran that simulation for various values of *k* (fig. 6). In this case, we see some benefits of decreasing *k* with large databases in conjunction with pronounced increases in the incorrect inclusion of differentiated noncausal loci. This is because, with 60 simulated populations, we are getting most of the causal loci even with $k = 0.5$.

FLAM versus Hypothesis Tests

After prescreening genes for differentiation in the twenty-population simulations, we did regressions of phenotype on allele frequency level at each locus, testing for slopes that were significantly different from zero. The *P*-values for each of these tests were saved and a Benjamini–Hochberg criteria was applied to determine which loci showed significant differences (see for instance Lovell et al. 2016). We did hypothesis testing with the twenty-population SNP simulations and linear allele frequency increases, using a 5% false discovery rate. With these simulated databases, hypothesis testing included 100% of causal loci in the sparse sets as well as 100% of the noncausal, differentiated loci. Thus, these hypothesis tests were *unable* to separate causal loci from differentiated noncausal loci.

Linkage

Genes that are tightly linked to causal loci would be expected to show allele frequency variation that closely follows the variation at the linked, causal locus. FLAM would be expected to include such tightly linked loci in the sparse lists. However, as the correlation in allele frequencies between the causal and noncausal linked locus decay, with increasing distance along a chromosome, we would anticipate that FLAM could start to differentiate such loci.

We created databases of 20 populations with varying levels of allele frequency correlations between the causal and

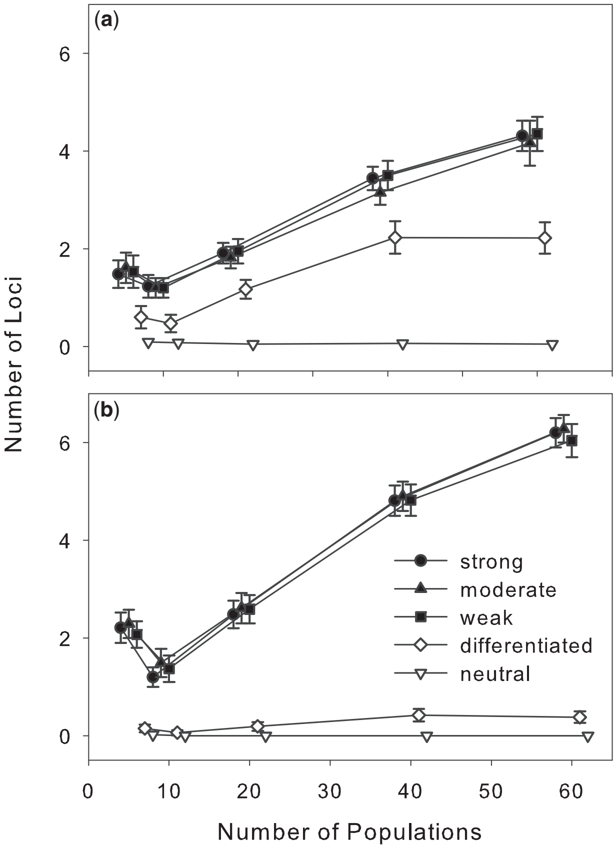


Fig. 2. The number of loci in the final sparse set for the causative loci and noncausative loci as a function of the total number of populations in the database. There are three categories of causative loci, strong, moderate and weak based on their phenotypic effects and two categories of noncausative loci, differentiated and neutral. The phenotypic standard deviation was varied as (*a*) 0.05 and (*b*) 0.005. The bars are 95% confidence intervals based on the 100 independent databases.

noncausal differentiated loci (see Materials and Methods). The simulation results (fig. 7) show a gradual decline in the inclusion of noncausal differentiated loci as the correlation with causal loci declines. The rate of inclusion drops to levels near those of independent loci at correlations produced by linkage disequilibrium of <0.40 (fig. 7). Estimates of linkage in laboratory populations of *Drosophila melanogaster* (Teotónio et al. 2009) suggest loci with correlations of 0.4 or less may encompass 10–100 kb. Accordingly, when analyzing data in the next section SNPs were partitioned into 50 kb regions.

Drosophila Data

Analysis of the genome-wide data structured in 50 kb SNP units yielded 194 SNP's. We first applied FLAM to a composite phenotype determined from a principal component analysis of pupal and adult development time, early and late fecundity, as well as adult mortality. We also applied FLAM to each phenotype individually. We applied BCD optimization to 100 permuted versions of the database. From these results, we assembled sparse lists based on $k = 0.2$, 0.3, 0.4, and 0.5. Using the six phenotypes with the 20 population database we found a total of 53 SNPs using at least one of the criteria (fig. 8). We identified 31 SNPs using the 30 population database for the pupal and adult development time phenotypes (fig. 9).

No SNP appears on the sparse lists of all phenotypes. However, there is a good deal of sharing between some phenotypes. For instance, pupal and adult development time share six SNPs in the 20-population database and seven in the 30-population database (figs. 8 and 9). But adult fecundity and adult mortality share only one SNP in common, even though between them there are a total of 23 SNPs in their sparse lists (fig. 8). Pupal development time in the 20 and 30 population databases share two SNPs in common. Adult development time in the 20 and 30 population databases share four SNPs in common. The 10 B populations have development times that are intermediate relative to the 10 A and 10 C populations. Thus, if FLAM included some noncausal differentiated loci in the 20 population analysis, these would be expected to be eliminated in the 30 population database analysis if their allele frequencies were not intermediate in the B populations.

We studied how well FLAM does at predicting phenotypes using the following method. Only loci from the $k = 0.5$ sparse list were used. We made training sets of 16 populations and predicted the phenotypes of the remaining four populations for the twenty-population databases. For the thirty-population databases, we trained with 8 A's, 8 B's and 8 C's, and then predicted the phenotypes of the remaining sample populations. We created all possible sets under the constraint that no population was used in the prediction set more than once. For the compound phenotype developed for the twenty-population data, the correlation between the predicted phenotype and the observed phenotypes was 0.96. In the twenty-population databases, pupal development time, adult development time, early fecundity, adult fecundity and adult mortality had predicted versus observed correlations of 0.94, 0.97, 0.82, 0.76, and 0.94, respectively. In the thirty-population database, the correlation for pupal development time was 0.90 and for adult development time 0.96.

Discussion

We have focused on a special kind of genetic material: experimentally evolved populations that have been exposed to different selection regimes for long periods of time and have consequently evolved large-scale genomic and phenotypic differentiation. Evidence from computer simulations has shown that FLAM can effectively sort out those loci that are differentiated and have a causal effect on a phenotype versus those that are differentiated but do not have a causal effect. The effectiveness of FLAM increases as 1) the number of independent populations increases, 2) the levels of within-treatment genetic variation increases, and 3) the level of non-genetic phenotypic variation between populations decreases.

The method will not identify all causative loci, even when the three conditions described above are relatively favorable. We also expect that any causative locus that does not show variation among replicate populations within selection regime will not be detected by FLAM—since it will have all the characteristics of a differentiated noncausal locus. In

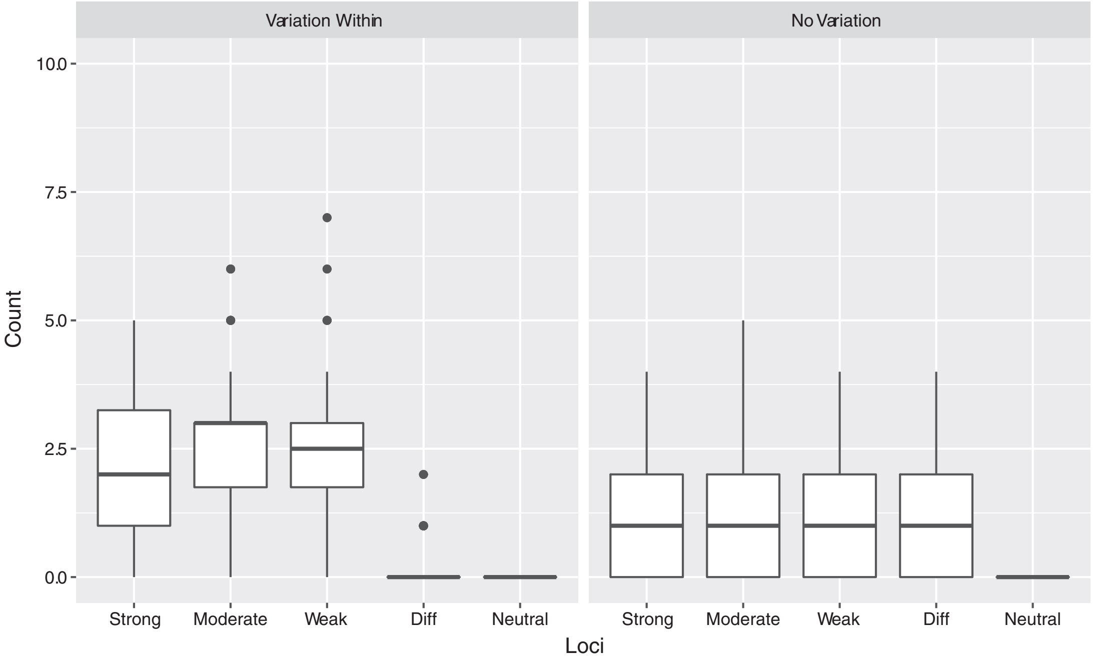


Fig. 3. The effects of subpopulation variation. Both panels show box plots of the number of loci in the final sparse sets among the 100 independent databases. The simulation shown in the right panel was based on a 20-population SNP database with ten populations that have the same high allele frequency and ten populations having the same low allele frequency. The left panel exhibits variation within the high and low groups. In the absence of replicate population variation, FLAM is unable to effectively sort out the noncausal differentiated loci from the causative loci.

particular, causal loci that become fixed in all replicate populations due to selection will be less likely to be detected. [However, no such loci are found in the experimental material that we have produced in our laboratory (e.g. Graves et al. 2017).] Likewise, loci that have a relatively weak effect on the phenotype and have causal patterns that are substantially different from those of loci that have strong phenotypic effects will not be detected by FLAM.

The most common setting for experimental evolution studies is to have a replicated set of control populations and a replicated set of experimental populations. However, if multiple selection regimes exist, resulting in intermediate phenotype values, we would expect this to facilitate FLAM's ability to sort out the causal from the differentiated noncausal loci. That is because it is less likely that the intermediate populations will follow the same sort of intermediate differentiation for noncausal loci. In effect, the intermediate populations are serving a similar role to replicate populations that show within-treatment phenotypic variation due to genetic differentiation.

Because of the large differentiation in the A and C *Drosophila* populations of Burke et al. (2016) and Graves et al. (2017) we studied here, FLAM does an exceptionally good job of predicting phenotypes in test populations not used to train the FLAM parameters. The correlation between predicted and observed phenotypes was typically well over 0.9. Speed and Balding (2014) applied their best linear unbiased prediction method to a variety of disease disorders in the Wellcome Trust Case Control Consortium; they obtained cross-validation correlations of around 0.3. Those data consisted of nearly 6,000 people. Makowsky et al. (2011), using similar models, could only explain ∼15–36% of the variation in human height, which has a heritability of ∼0.8. We believe the superior predictive power of FLAM applied to laboratory selected populations is due to the very large phenotypic differentiation that can be achieved by experimental evolution.

Studies of geographically differentiated populations of *Drosophila* and laboratory selected populations have used a variety of standard statistical tests and regression analyses to infer causal loci (Bochdanovits et al. 2003; Griffin et al. 2017). None of these studies has looked seriously at how effective these methods are. As we show in this study, standard regression analysis, even when it uses methods to control for false discovery, does a very poor job of differentiating causal from noncausal loci.

Materials and Methods

Simulated SNP Frequencies

The simulated populations are characterized by SNP frequencies at genetic markers that are directly responsible for a complex phenotype, here called *causative loci*, as well as loci that do not affect the phenotype, called *noncausative* loci. We further define three subtypes of causative loci: those with strong, moderate, or weak effects on the phenotype. There are also two subtypes of noncausative loci for a particular phenotype: those that show genetic differentiation between selection regimes and those that do not. The differentiated but noncausative loci are expected to exist in laboratory selected populations, because it is difficult to contrive experimental evolution paradigms that only select on a single phenotype. Such differentiated noncausative loci present a key challenge to statistical techniques that are used to

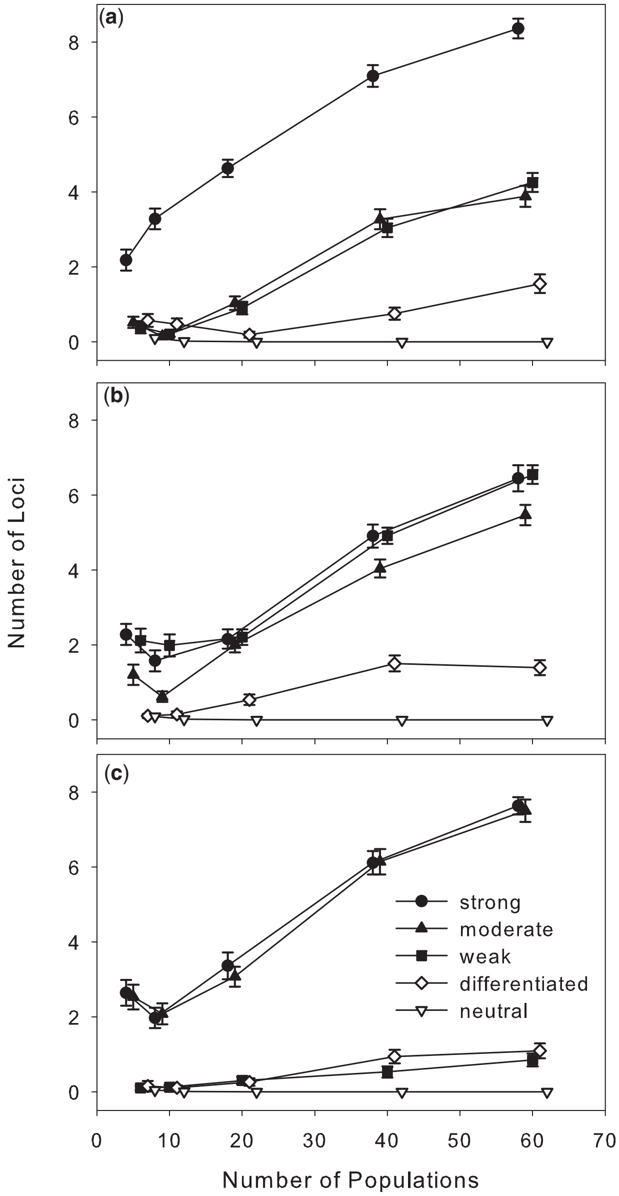


Fig. 4. The number of loci in the final sparse set for the causative loci and noncausative loci as a function of the total number of populations in the database. The phenotypic standard deviation was 0.005 in all cases. The plateau patterns of allele frequency variation (fig. 1*c* and *d*) were used for the (*a*) strong causal loci, (*b*) moderate causal loci, and (*c*) weak causal loci.

infer which loci are causative for a particular type of phenotypic differentiation.

For each population-*i* ($i = 1, \ldots, n$), we let the population allele frequency for the three categories of causative loci at locus-*j* ($j = 1 \ldots n_s$, n_m, or n_w) be $p_{i,\,j}$. Coverage over populations and loci varies, so for each locus we treat coverage as a random variable, $\hat{N}_{i,j}$, which we sample from a normal distribution with a mean of 76 and standard deviation of 16 based on our observations in the 10 A and C-type populations (described later). To avoid extremely small values of $\hat{N}_{i,j}$, we truncate the lower end of the distribution to 29, the lowest average coverage observed in any of the populations. To generate a synthetic database, we then chose sample allele

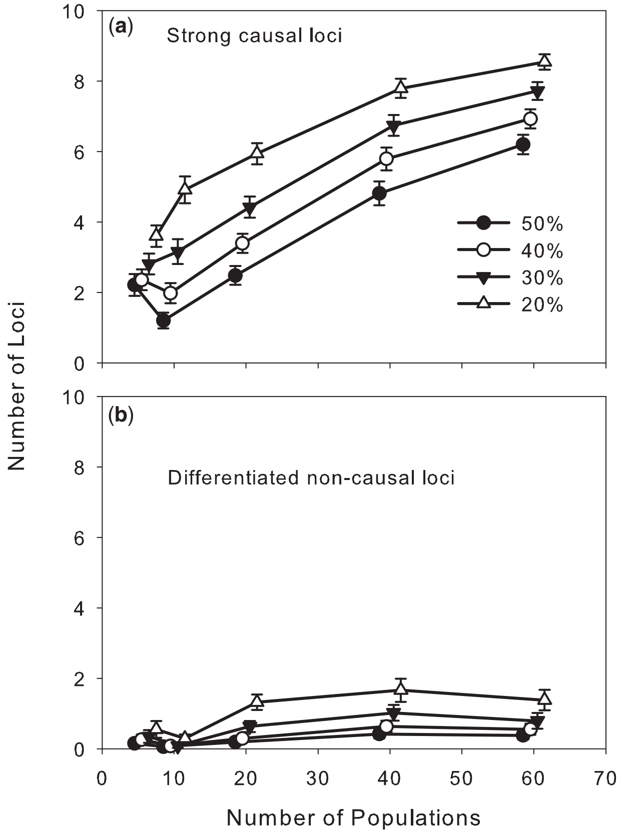


Fig. 5. The number of loci in the final sparse set for the (*a*) strong causative loci and (*b*) noncausative loci as a function of the total number of populations in the database and various values of *k* (expressed as a percent). These results were based on the simulations run in figure 2*b*.

frequencies, $\hat{p}_{i,s}$ ($s = 1 \ldots n_s + n_m + n_w$), at the ten strong, moderate, and weak loci as $B(\hat{N}_{i,j}, p_{i,j})$, where j= (s - 1 modulo 10) + 1 (when $n_s = n_m = n_w = 10$).

The phenotype of the population depends on the population allele frequencies, $p_{i,\,j}$, not the sample allele frequencies. Thus, the phenotype (P_i) in population-*i* is,

$$P_i = \left(\sum_{s\in n_s} a_s p_{i,s} + \sum_{s\in n_m} a_m p_{i,s} + \sum_{s\in n_w} a_w p_{i,s}\right) \left(a_s n_s + a_m n_m + a_w n_w\right)^{-1} + \varepsilon_{ij}$$

where ε_{ij} is a measure of environmental noise that was assumed to have a normal distribution with a mean of 0 and standard deviation of 0.005. The effects of these loci on the phenotype were set as a_s (=1), a_m (=0.5), and a_w (=0.1) for strong, moderate and weak loci respectively. We set $n_s = n_m = n_w = 10$.

The total number of populations in these simulations varied from 6 to 60 (fig. 1*a* and *b*). In one set of simulations, allele frequencies at the causative loci were assumed to increase linearly over the subpopulations (fig. 1*a* and *b*). In a second set of simulations, we allowed allele frequencies in either the strong, moderate, or weak causative loci to have a plateau-shaped change in allele frequencies (fig. 1*c* and *d*). The noncausative differentiated loci showed the same range of allele

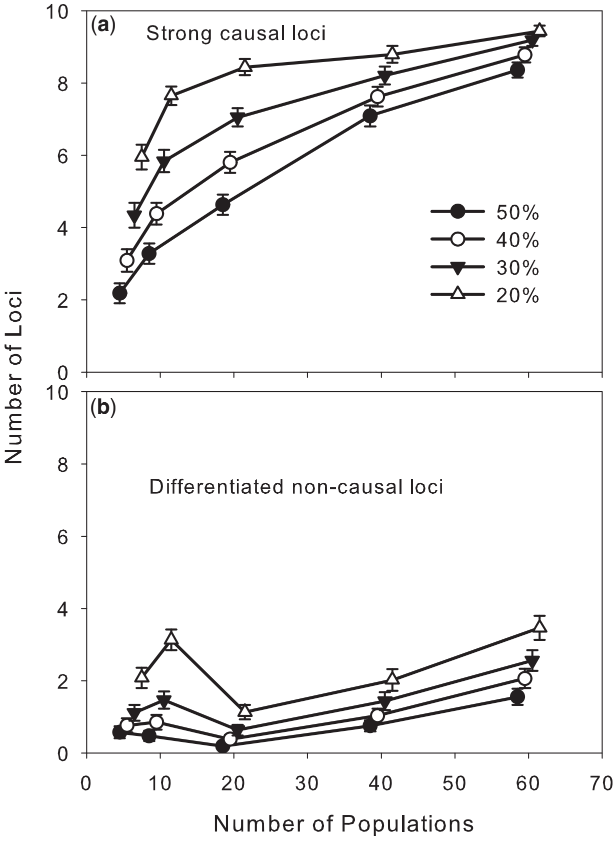


FIG. 6. The number of loci in the final sparse set for the (*a*) strong causative loci and (*b*) noncausative loci as a function of the total number of populations in the database and various values of *k* (expressed as a percent). These results were based on the simulations run in figure 4*a*.

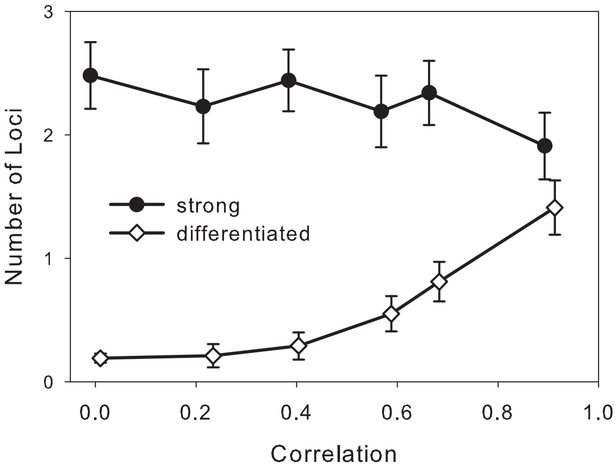
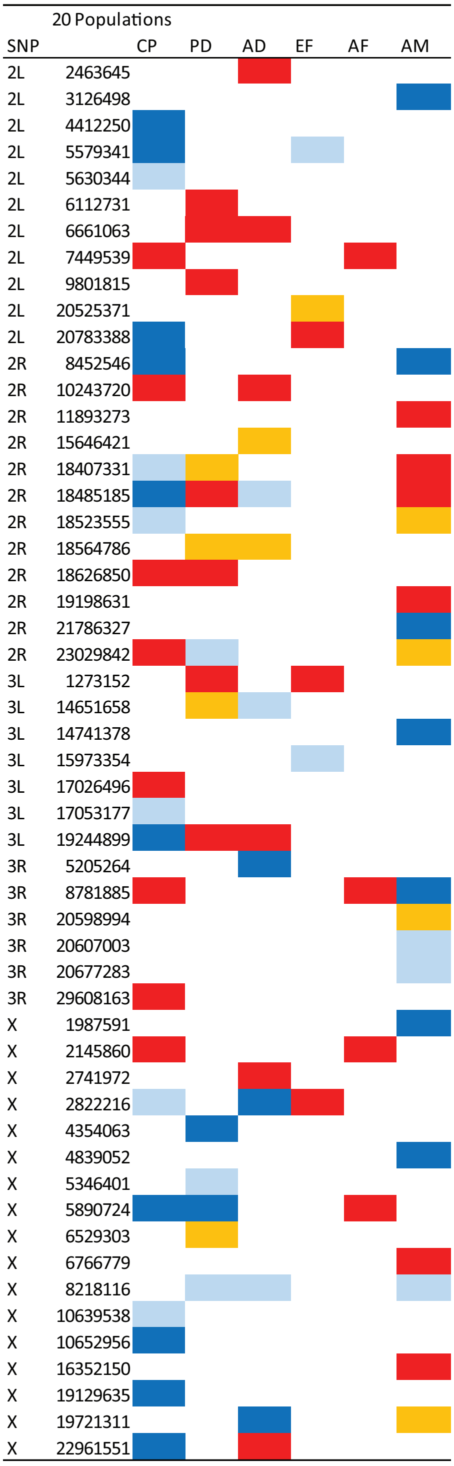


FIG. 7. The number of loci in the final sparse set for the strong causative loci (solid circle) and the differentiated, noncausative loci (open diamond) as a function of correlation (linkage) between these loci.



20 Populations							
SNP		CP	PD	AD	EF	AF	AM
2L	2463645			🟥			
2L	3126498						🟦
2L	4412250	🟦					
2L	5579341	🟦			◻		
2L	5630344	◻					
2L	6112731		🟥				
2L	6661063		🟥	🟥			
2L	7449539	🟥				🟥	
2L	9801815		🟥				
2L	20525371				🟧		
2L	20783388	🟦			🟥		
2R	8452546	🟦					🟦
2R	10243720	🟥		🟥			
2R	11893273						🟥
2R	15646421			🟧			
2R	18407331	◻	🟧				🟥
2R	18485185	🟦	🟥	◻			🟥
2R	18523555	◻					🟧
2R	18564786		🟧	🟧			
2R	18626850	🟥	🟥				
2R	19198631						🟥
2R	21786327						🟦
2R	23029842	🟥	◻				🟧
3L	1273152		🟥		🟥		
3L	14651658		🟧	◻			
3L	14741378						🟦
3L	15973354				◻		
3L	17026496	🟥					
3L	17053177	◻					
3L	19244899	🟦	🟥	🟥			
3R	5205264			🟦			
3R	8781885	🟥				🟥	🟦
3R	20598994						🟧
3R	20607003						◻
3R	20677283						◻
3R	29608163	🟥					
X	1987591						🟦
X	2145860	🟥				🟥	
X	2741972			🟥			
X	2822216	◻		🟦	🟥		
X	4354063		🟦				
X	4839052						🟦
X	5346401		◻				
X	5890724	🟦	🟦			🟥	
X	6529303		🟧				
X	6766779						🟥
X	8218116		◻	◻			◻
X	10639538	◻					
X	10652956	🟦					
X	16352150						🟥
X	19129635	🟦					
X	19721311			🟦			🟧
X	22961551	🟦		🟥			

FIG. 8. The sparse lists for the 20 A and C *Drosophila* populations with SNP's at 50 kb intervals. The phenotypes are compound phenotype (CP), pupal development (PD), adult development (AD), early fecundity (EF), adult fecundity (AF), and adult mortality (AM). The color codes indicate which criteria the SNP has satisfied: red ($k = 0.5$), red and orange ($k = 0.4$), red, orange, and light blue ($k = 0.3$) and red, orange, light blue and dark blue ($k = 0.2$).

frequencies as the causative loci, however the subpopulation designations were randomly shuffled. As an example, in the 20 population simulations the population allele frequencies at the causative loci in the 10 populations with low phenotypic values were, $\boldsymbol{p}_C = (0.45, 0.46, 0.47, 0.48, 0.49, 0.50, 0.51, 0.52, 0.53, 0.54)$. Allele frequencies at the noncausative loci, $\boldsymbol{p}_{NC}$, were based on a random sample without replacement from $\boldsymbol{p}_C$. Thus, the correlation between $\boldsymbol{p}_C$ and $\boldsymbol{p}_{NC}$ would be 0.

30 populations			
SNP		PD	AD
2L	4265950	🟥	
2L	4331882		🟥
2L	5894963		🟦
2L	5994543	🟧	🟦
2L	6090933		🟥
2L	6112731		🟥
2L	9801815		🟦
2L	20801145		🟦
2R	8569937	🟧	🟥
2R	11846148	🔹	
2R	11932032		🔹
2R	18626850	🟦	
2R	21786327	🟥	🟥
3L	684844		🔹
3L	719871		🟦
3L	774317		🟥
3L	7315037	🟦	
3L	14651658	🟥	🟥
3L	19244899		🟧
3L	21009121	🟥	
3R	5205264	🟥	🟥
3R	14139656	🟦	🟦
3R	21768852	🟥	
3R	22114121		🟦
3R	29608163	🟥	🟥
X	1987591		🔹
X	2073341		🟥
X	6429797		🟦
X	8218116		🟦
X	16270729		🔹
X	19721311	🟥	

Fig. 9. The sparse lists for the 30 A, B, and C *Drosophila* populations. The color codes are the same as figure 8.

For two bi-allelic linked loci, with the most common allele having frequencies of p_1 and q_1, the allele frequency correlation between these two loci can be represented as $D[p_1(1 - p_1)q_1(1 - q_1)]^{-1/2}$ where D is the linkage disequilibrium coefficient. Thus, to study the effects of linkage we manipulated the correlation between $\boldsymbol{p}_C$ and $\boldsymbol{p}_{NC}$. The allele frequency vectors (table 1) retained the same allele frequencies as the causative loci, but the arrangement among the subpopulations varied with the specific distributions shown in table 1.

The phenotypic differentiation of the simulated high-phenotype and low-phenotype populations using a sample of 20 populations was substantial (fig. 10), but not unlike the levels we see in our laboratory populations of *Drosophila* (vid. Burke et al. 2016). For the other simulations, the phenotypic variation was proportional to the mean allele frequencies. The mean phenotype within each population showed low levels of variation due to sampling effects (fig. 11). The simulation also included 10 differentiated, noncausative loci. The allele frequencies at these loci were the same as those of the causative loci, although the subpopulation variation was uncorrelated with the variation at the causative loci. There were an additional 1,960 noncausative loci with a mean allele frequency of 0.9.

Statistical Inference with FLAM

In both our simulated and real-world data, we assume that the SNP frequency across the m loci has been measured in n independent populations along with n-phenotypes, $\mathbf{P} = (P_1, P_2, \ldots, P_n)$. At locus-$j$, for instance, we assume that the allele frequencies can be ordered as, $\hat{p}_{1j} < \hat{p}_{2j} < \ldots < \hat{p}_{nj}$. The regression relationship, $E[P_i|\hat{p}_{ij}] = \theta_i$, of the FLAM (Petersen et al. 2016) will yield estimates of the parameter vector, $\boldsymbol{\theta}_j = (\theta_{1j}, \ldots, \theta_{nj})$ subject to,

$$\underset{\theta_j \in \mathbb{R}^n}{\mathit{minimize}} \ \frac{1}{2}\left\|\mathbf{P} - \theta_j\right\|_2^2 + \lambda\left\|\mathbf{D}\theta_j\right\|_1 \qquad (1)$$

where

$$\mathbf{D} = \begin{pmatrix} 1 & -1 & 0 & \ldots.0 \\ 0 & 1 & -1 & \ldots.0 \\ . & . & . & . \\ 0 & 0\ldots. & 1 & -1 \end{pmatrix}$$

(Petersen et al. 2016).

Large values of the tuning parameter λ will tend to make $|\theta_{i\text{-}1,\,j} - \theta_{i,\,j}|$ equal to zero. Hence, the final function will be a series of steps with jumps or knots that are adaptively chosen.

Over all loci, we add to the optimization problem in equation (1) a group lasso penalty function that will encourage whole $\boldsymbol{\theta}_j$ vectors to be zero and thus serve to eliminate uninformative loci yielding,

$$\underset{\theta_0 \in \mathbb{R},\, \theta_j \in \mathbb{R}^n,\ 1 \leqslant j \leqslant m}{\mathit{minimize}} \ 1/2\left\|\mathbf{P} - \sum\nolimits_{j=1}^{m} \theta_j - \theta_0 1\right\|_2^2 \qquad (2)$$
$$+ \alpha\lambda \sum\nolimits_{j=1}^{m} \left\|\mathbf{D}\mathbf{M}_j\theta_j\right\|_1 + (1 - \alpha)\lambda \sum\nolimits_{j=1}^{m} \left\|\theta_j\right\|_2,$$

where $\mathbf{M}_j$ is a matrix that orders the values of $\hat{\boldsymbol{p}}_j$ from smallest to largest. Equation (2) adds a second tuning parameter α, which ranges from 0 to 1 (Petersen et al. 2016). The R-function, *flamCV* (in the *flam* package), will search for the best λ based on the cross-validation error rates for a given value of α. In our analyses, we used a grid of 19 α-values to find the best model ($\alpha = 0.05, 0.1, \ldots, 0.95$).

The solutions to (2) are characterized by a sparse set of loci where $\left\|\theta_j\right\|_2 \neq 0$. Although there is a global minimum for the objective function (2), there is not a unique solution. One method for finding this minimum is called block coordinate descent ("BCD," Friedman et al. 2007). While it is fast and robust compared with other methods, the sparse set of loci that are identified by BCD depend on the arrangements of loci in the matrix of independent variables.

Table 1. The Allele Frequency Vectors Used to Study the Effects of Linkage in Twenty Population Simulations.

Loci	Correlation	Phenotype	Allele Frequencies
Causative		**Low**	**(0.45, 0.46, 0.47, 0.48, 0.49, 0.5, 0.51, 0.52, 0.53, 0.54)**
		High	**(0.85, 0.86, 0.87, 0.88, 0.89, 0.9, 0.91, 0.92, 0.93, 0.94)**
Noncausative	**0.224**	**Low**	**(0.49, 0.5, 0.51, 0.52, 0.45, 0.46, 0.47, 0.48, 0.53, 0.54)**
		High	**(0.89, 0.9, 0.91, 0.92, 0.85, 0.86, 0.87, 0.88, 0.93, 0.94)**
Noncausative	**0.394**	**Low**	**(0.49, 0.5, 0.52, 0.45, 0.46, 0.48, 0.47, 0.53, 0.51, 0.54)**
		High	**(0.89, 0.9, 0.92, 0.85, 0.86, 0.88, 0.87, 0.93, 0.91, 0.94)**
Noncausative	**0.578**	**Low**	**(0.48, 0.49, 0.5, 0.45, 0.46, 0.47, 0.53, 0.54, 0.51, 0.52)**
		High	**(0.88, 0.89, 0.9, 0.85, 0.86, 0.87, 0.93, 0.94, 0.91, 0.92)**
Noncausative	**0.673**	**Low**	**(0.48, 0.49, 0.5, 0.45, 0.46, 0.47, 0.51, 0.52, 0.53, 0.54)**
		High	**(0.88, 0.89, 0.9, 0.85, 0.86, 0.87, 0.91, 0.92, 0.93, 0.94)**
Noncausative	**0.903**	**Low**	**(0.47, 0.48, 0.45, 0.46, 0.49, 0.5, 0.51, 0.52, 0.53, 0.54)**
		High	**(0.87, 0.88, 0.85, 0.86, 0.89, 0.9, 0.91, 0.92, 0.93, 0.94)**

The correlation column reports the correlation coefficient between the noncausative loci and the causative loci in the same phenotype class.

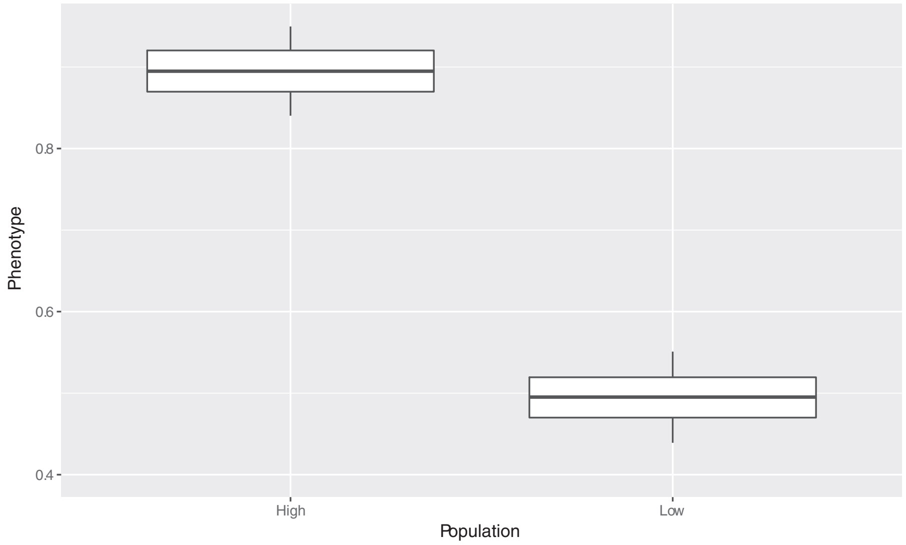


Fig. 10. Box plots of the phenotypic distribution in the 10 high and 10 low populations used in the 20 population simulations. Phenotypes are based on SNP frequencies.

We identified causative loci with the following algorithm. Find a solution to (2) using BCD and save the sparse set. Then permute the columns of the matrix of independent variables and solve (2) again. Repeat the permutation and solution steps 100 times. Enumerate the frequency of occurrence of each SNP among these 100 sparse sets. Let the frequency of the most common SNP among the 100 set be C_{max}. Identify as the causative loci only those that occur greater than kC_{max}, where $k = 0.2, 0.3, 0.4,$ or 0.5. We present evidence supporting this rule in the Results section.

SNP Data

Read Mapping

For this analysis, we used the genome-wide SNP data previously published in Graves et al. (2017). This data set contains pooled Illumina paired-end sequence data from 30 experimentally evolved *D. melanogaster* populations: $ACO_{1–5}$, $AO_{1–5}$, $B_{1–5}$, $BO_{1–5}$, $CO_{1–5}$, and $nCO_{1–5}$. (See Graves et al. 2017 for extraction details.) Raw fastq files were obtained and mapped to the *D. melanogaster* reference genome (version 6.14) using bwa mem (BWA version 0.7.8) with default settings (Li and Durbin 2009). The resulting SAM files were filtered, sorted, and converted to BAM files using the view and sort commands in SAMtools (Li et al. 2009). We only selected reads mapped in proper pairs with a minimum mapping quality of 20. The rmdup command in SAMtools was then used to remove potential PCR duplicates. As each population in Graves et al. (2017) was sequenced twice, there were two bam files corresponding to each population at this stage. BAMtools was used to combine pairs of BAM files corresponding to the same populations. The 30 resulting BAM files were then combined into a single mpileup file using SAMtools. This mpileup file was then further converted to a "synchronized" file, a format that contains allele counts for all bases in the reference genome and for all populations being analyzed, using the PoPoolation2 software package (Kofler et al. 2011). Lastly, RepeatMasker 4.0.3 (http://www.repeatmasker.org) was used to create a gff of high repetitive regions found in the 6.14 release of the *D. melanogaster* reference genome. These regions were then removed from our sync file once again using PoPoolation2.

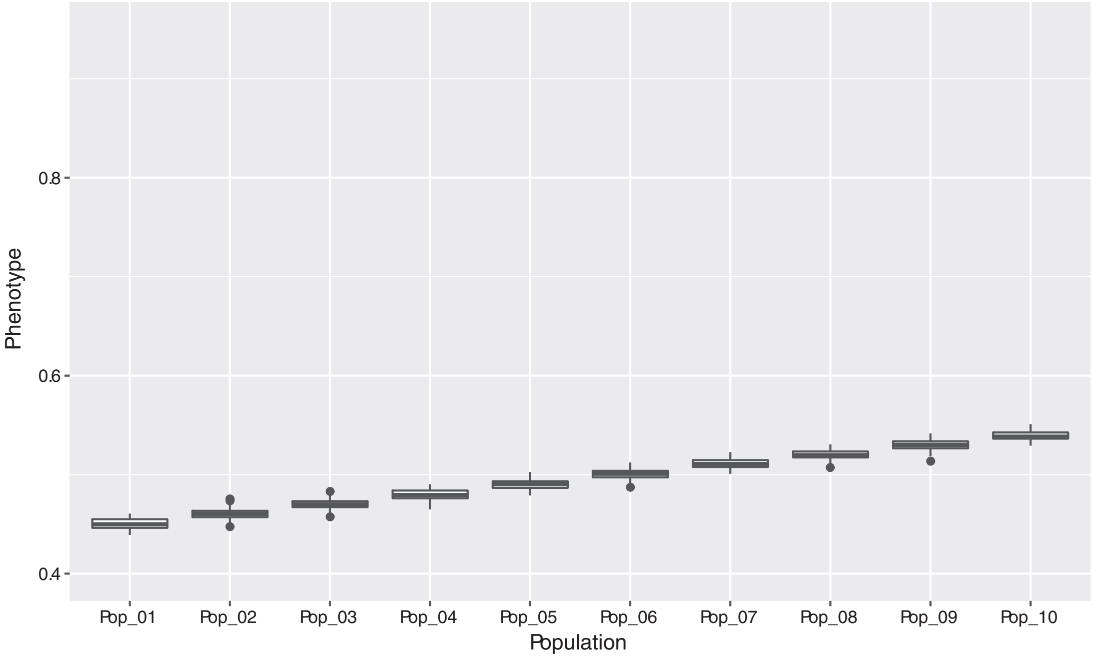


Fig. 11. Box plots of phenotypic variation in the 10 subpopulations within the low phenotype group shown in figure 10 due to allele frequency sampling variation and environmental variation.

Identifying Candidate Regions of the Genome

We only considered biallelic sites and required each site to have coverages between 20× and 200× in each of the 30 populations. We also required each site to have a minimum minor allele frequency of 2% across all 30 populations. All sites failing to meet these criteria were discarded. To test for SNP differentiation, we used the Cochran–Mantel–Haenzel (CMH) test as implemented in PoPoolation2. CMH tests were performed between the 10 A-type populations ($ACO_{1–5}$ and $AO_{1–5}$) and 10 C-type populations ($CO_{1–5}$ and $nCO_{1–5}$) at all sites meeting our SNP calling criteria. Populations were paired based on treatment and replicate number (e.g. ACO_1 was paired with CO_1, AO_1 with nCO_1, etc.). To correct for multiple comparisons, genetic drift, and sampling, we used the permutation approach featured in Graves et al. (2017). Briefly, we randomly assigned population to one of two groups, and then performed CMH tests at each polymorphic site in the shuffled data set to generate null distributions of *P*-values. We did this 1,000 times, and each time we recorded the smallest *P*-value generated. We then used the quantile function in R to establish a significance threshold that defines the genome-wide false-positive rate, per site, at 5%. This process resulted in a significance threshold of 1.95×10^{-212}. Using this significance threshold, we identified a total of 4,211 candidate SNPs between the A-type and C-type populations spread out across the five major chromosome arms.

Next, we identified 50 kb regions based on our list of 4,211 candidate SNPs. For the 50 kb candidate regions, we first divided each chromosome arm into 50 kb windows. All windows containing <3 candidate SNPs were discarded. We then went through each of the remaining windows, and recorded the position in each window with the smallest *P*-value from our CMH tests. This resulted in a list of 194 positions representing the 194 50 kb windows that met our criteria. These positions and their associated SNP frequencies were then used as inputs in our FLAM analysis.

Phenotype Data

Phenotypes were measured on individuals from 20 populations that varied by their age-of-reproduction. Ten populations were reproduced at 10 days of life (measured from egg). Five of these populations are called ACO and five AO. Their historical relationship is summarized in Burke et al. (2016). Another ten populations were reproduced at 28 days: five are called CO and five *n*CO (Burke et al. 2016). We used five different phenotypes that show substantial differentiation between the A and C type populations: egg-to-pupa development time, egg-to-adult development time, fecundity at 264 h of life (measured from egg), fecundity at adult age 27 days and adult mortality at age 16 days (Burke et al. 2016). We combined these five phenotypes into a single compound phenotype using the first principal component derived from centered and scaled values of the five phenotypes and the R program *prcomp* (R Core Team 2015).

We also analyzed a group of 30 populations that included the 10 A, 10 C, and 10 B populations. The B populations reproduce at day 14 of life. They show pupal and adult development times that are intermediate to the A and C populations. FLAM was used with these 30 populations to infer causal SNPs for pupal and adult development times.

Acknowledgments

We thank Ashley Petersen for helpful discussions about FLAM and three anonymous referees for comments on the manuscript. This research has been supported by the Ayala School of Biological Sciences, University of California, Irvine.

References

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. 57:289–300.

Bochdanovits Z, van der Klis H, de Jong G. 2003. Covariation of larval gene expression and adult body size in natural populations of *Drosophila melanogaster*. *Mol Biol Evol*. 20:1760–1766.

Burke MK, Barter TT, Cabral LG, Kezos JN, Phillips MA, Rutledge GA, Phung KH, Chen RH, Nguyen HD, Mueller LD, et al. 2016. Rapid divergence and convergence of life-history in experimentally evolved *Drosophila melanogaster*. *Evolution* 70:2085–2098.

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193:327–345.

Friedman J, Hastie T, Hofling H, Tibshirani RT. 2007. Pairwise coordinate optimization. *Ann Appl Stat*. 1:302–332.

Garland T, Rose MR. 2009. Experimental Evolution. Berkeley (CA): University of California Press.

Graves JL, Hertweck KL, Phillips MA, Han MV, Cabral LG, Barter TT, Greer LF, Burke MK, Mueller LD, Rose MR. 2017. Genomics of parallel experimental evolution in *Drosophila*. *Mol Biol Evol*. 34:831–842.

Griffin PC, Hangartner SB, Fournier-Level A, Hoffmann A. 2017. Genomic trajectories to desiccation resistance: convergence and divergence among replicate selected *Drosophila* lines. *Genetics*. 205:871–890.

Kofler R, Pandey RV, Schlötterer C. 2011. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* 27:3435–3436.

Lander ES, Botstein D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121:185–199.

Landis JR, Heyman ER, Koch GG. 1978. Average partial association in three-way contingency tables: a review and discussion of alternative tests. *Int Stat Rev*. 46:237–254.

Lewontin RC. 1974. The Genetic Basis of Evolutionary Change. New York: Columbia University Press.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.

Lovell JT, Shakirov EV, Schwartz S, Lowry DB, Aspinwall MJ, Taylor SH, Bonnette J, Palacio-Mejia JD, Hawkes CV, Fay PA, et al. 2016. Promises and challenges of eco-physiological genomics in the field: tests of drought responses in switchgrass. *Plant Physiol* doi:10.1104/pp.16.00545.

Makowsky R, Pajewshi NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de los Campos G. 2011. Beyond missing heritability: prediction of complex traits. *PLoS Genet*. 7:e1002051.

Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829.

Petersen A, Witten D, Simon N. 2016. Fused lasso additive model. *J Comp Graph Stat*. 25:1005–1025.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: https://www.R-project.org/.

Speed D, Balding DJ. 2014. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res*. 24:1550–1557.

Teotónio H, Chelo IM, Bradić M, Rose MR, Long AD. 2009. Experimental evolution reveals natural selection on standing genetic variation. *Nat Genet*. 41:251–257.

Weissbrod O, Geiger D, Rosset S. 2016. Multikernal linear mixed models for complex phenotype prediction. *Genome Res*. 26:969–979.