**File S1: Supplementary Methods and Results**

*Candidate SNP identification*

While a Fisher’s exact test is commonly used in Poolseq studies to test for differentiation in allele frequencies (Fabian *et al.* 2012; Fischer *et al.* 2013; Kapun *et al.* 2014), some authors have suggested this rather simple approach is not appropriate for pooled data (Lynch *et al.* 2014). As a second step, we also used the approach suggested by Lynch *et al.* (2014) that takes into account sequencing error rate and pool size in maximum likelihood estimation of allele frequency within a pool. We implemented the test suggested by Lynch *et al.* (Formula 8) in R, as a second approach to testing whether mass-bred and generation 13 pools differed in allele frequency at each SNP:

where and are the likelihood of the reads observed in pool 1 and 2 respectively, under a situation where they come from a total population with homogeneous allele frequency; and and are the likelihoods of the reads observed in pool 1 and 2 respectively, where pool allele frequencies are actually heterogeneous. (See below, and Lynch *et al.* (2014), for further details on the likelihood calculations). The resulting likelihood ratio was subjected to a Χ2 test with 1 degree of freedom. As before, we applied a Bonferroni correction to α = 1×10-6 to adjust for multiple comparisons. This correction was calculated for each replicate in order to allow for the differences in total number of SNPs called.

Our third step was to ensure we were focusing only on loci that had diverged from their initial frequency due to selection, not to drift over the 13 generations of the experiment. To address this, expected drift was simulated 1000 times per starting minor allele frequency, for each starting frequency value from 0.01 to 0.50 by steps of 0.01. A zero-truncated model was run for 13 generations, with a constant population size of 200 diploid individuals (the estimated census size at the start of each generation) for the nuclear genome and 200 haploid individuals for the mitochondrial genome. The upper and lower bounds of the resulting 95% confidence interval describing the final frequency were retained. For loci with starting allele frequency 0.01, the 95% confidence interval for 0.02 was used as a conservative estimate of drift expectations given the limited pool size.

For each locus, the overall major and minor allele were identified based on read counts (major allele defined as major in the mass-bred pool). A ML model was set up following Lynch *et al.* Formula 7, as described in the text:

*pM* = ML estimate of the major-allele frequency

2N = haploid pool size

*𝜖* = sequencing error rate (conservatively set at 0.01)

The model was implemented in R using the mle2 function in the bbmle package (BolkerR Development Core Team 2014), run using the ‘optimize’ optimizer, with the point ML estimate as the starting probability and a vector of (0,1) calls representing the observed allele counts for that locus as the input data. The ‘confint’ function was then used to summarise the 95% CI around the ML frequency estimate. CI intervals were calculated in this way for the initial (mass-bred) and final (generation 13) allele frequency for each SNP.

As an additional, conservative method of assessing the significance of frequency change, for loci showing an increase in MAF between the initial and final observations, the upper bound of the initial-observation CI was used as the ‘initial frequency’. The matching drift simulation was chosen; if the lower bound of the final-observation CI fell outside the 95% CI of the drift simulation, the frequency change was considered to be significantly outside that expected under drift alone. For loci that decreased in MAF, the lower bound of the initial-observation CI and the upper bound of the final-observation CI were used in the same way. If the MLE model did not converge, the observed MAF was used instead of the CI limit.

The script lynch\_pop\_comparison.R used to implement these tests is available at https://github.com/griffinp/desiccation-selection/

*Impact of coverage depth on SNP candidate status*

As sequencing coverage depth varied across the genome and among replicates (Table S1), we wanted to check whether coverage depth variation was contributing to the classification of SNPs as significantly differentiated between the mass-bred and generation 13 pools. We extracted coverage (total number of reads mapped to the location) for the candidate SNPs for each desiccation replicate. We also extracted coverage for a random set of 100,000 SNPs. For each desiccation replicate, we performed a one-sided t-test to ascertain whether the candidate SNPs showed significantly lower coverage than the random SNPs.

*Statistical power of detecting associations*

We performed 2×106 forward simulations based on the population and genomic context of this study to assess the statistical power of detecting associations. A focal gene carrying a single causal SNP with additive effect explaining either 0.01 or 0.001 of the trait broad-sense heritability was randomly chosen from the gene set annotated on chromosome 2L, 2R, 3L or 3R. The haplotype included the focal gene, a randomly-selected nearby gene ~10Kbp away and a distant gene ~1Mbp away to test the propensity for short- and long-distance linkage disequilibrium to generate false positive associations. For each gene trio, the same number of SNPs was set as in the real data. In order to conservatively model pre-existing linkage among allele frequencies, each SNP was assigned an allele drawn from a binomial distribution corresponding to the experimental frequency distribution constrained by covariance with neighbouring loci. The covariance matrix among SNPs was defined such that linkage within a gene followed a uniform distribution U(0, 0.8); linkage between genes located ~10Kbp apart followed U(0, 0.4); and linkage between genes located ~1Mbp apart followed U(0, 0.2). This approach defined each haplotype under a conservative model of background linkage, allowing higher linkage than that estimated for the set of real inbred lines in the Drosophila Genetic Reference Panel (Garud *et al.* 2015). The evolution (number of haplotypes and selection intensity) of the 400 initial haplotypes over the 13 generations of the model was then simulated to match the experimental design, assuming random mating. When simulating gametes from each parental haplotype, recombination rate across the genome was set according to the empirical measurements of Comeron *et al.* (2012). After 13 generations, the change in frequency for all SNPs of the gene trio was tested using a Fisher Exact Test with an error risk of 3.5 × 10-13, in line with the threshold used for the experimental candidate SNP identification (Table S1). Two million simulations were run. The significant associations detected in the gene bearing the causal variant were interpreted as true positives. Significant associations in the 10Kbp distant gene or the 1Mbp distant gene were interpreted as short-range or long-range false positives respectively. The script Simul\_SelectionRegime.py used to implement these tests is available at https://doi.org/10.5281/zenodo.166392 (Griffin 2016b).

*Impact of multiple-mapping SNPs on gene-level overlap among replicates*

We divided the candidate genes into single-mapped and multiple-mapped sets for the purpose of testing the impact of multiple-mapping on gene sharing. For each replicate, we extracted a list of candidate genes contributed to by single-mapping SNPs only, and a list of candidate genes contributed to by multiple-mapping SNPs only. Single-mapping SNPs were defined as SNPs that mapped to within 1000 bp of only one gene; multiple-mapping SNPs were defined as SNPs that mapped to within 1000 bp of two or more genes.

We examined the overlap among replicates in the resulting gene lists by counting the number of genes shared among each possible pair, trio, quartet and all five replicates (as was done for the full analysis). The overlap was counted for the single-mapped gene lists and for the multiple-mapped gene lists separately. If multi-mapping SNPs were driving patterns of gene overlap (i.e. increasing the estimated overlap among genes), we expected to see the slope of the single-mapped vs. multi-mapped gene overlap relationship to be much less than 1. In this situation, the categories (pairs, trios etc.) with higher total overlap should be over-represented in multiple-mapped genes. Conversely, if multi-mapping were having no effect on the level of total gene overlap, we would expect the slope of the single-mapped vs. multi-mapped gene overlap relationship to be close to 1.

**Results**

*Justification of experimental design and replicate line approach*

Experimental conditions that reduce the effective recombination rate can elicit haplotype-based selection that may not be relevant to natural populations (Tobler *et al.* 2013; Baldwin-Brown *et al.* 2014; Franssen *et al.* 2014). We believe we largely avoided this problem in this study, by starting with a large, diverse and well-mixed population, and maintaining a high effective population size over the course of the experiment (Figure 3), as recommended by Franssen *et al.* (2014) and Tobler *et al.* (2013). Also, Burke et al. (2014) noted apparent selection effects on a large proportion of the genome can be observed as an artifact when few replicates and timepoints are examined. Our study is necessarily limited to five replicates and two timepoints; however this is a higher level of replication than used in studies that have successfully identified known candidate loci for simple traits (e.g. Bastide *et al.* 2013). Baldwin-Brown *et al.* (2014) found that reasonable power to detect causative SNPs was obtained using ‘achievable’ experimental designs (no. replicate lines >= 5, no. starting haplotypes >= 32, no. diploid individuals >= 500, no. generations = 100, selection coefficient > 0.2). However this study addressed only the case of a constant selection coefficient, and only counted SNPs as significant if they were found to act across multiple replicate selected lines. The authors also acknowledged that their method did not span all possible parameter values and that (for example) loci with higher selection coefficients could potentially be identified using even much smaller studies. Another simulation study took the trait architecture into account using a QTL-like simulation design (Kessner and Novembre 2015) and examined the case of truncating selection, which is more relevant to our study. Over the course of a truncating-selection experiment, the mean realized selection coefficient *s* varies depending on the effect size of the QTL. Large-effect QTLs were shown to fix very quickly (most with effect size 0.3 fixed in under 20 generations from a starting frequency of 0.25). Related to this point, the relationship between the number of generations and the power of the study is actually rather complex, as these authors showed that most QTLs differentiate relatively early under strong selection (e.g. by generation 20) and after this time drift is likely to drive many neutral variants to fixation. By examining evolution over 13 generations and replicating the control as well as selected lines, we should have reasonable power to detect causative SNPs of large effect size. However, some indications remained of a possible relationship between sequencing coverage depth and candidate SNP number (Table S1), despite stringent and consistent multiple-testing corrections.

In this study, control-line replication allowed us to identify and exclude loci potentially involved in lab adaptation, and we retained only loci that showed an allele frequency change greater than that predicted under genetic drift. Although experimental limitations remain that have likely resulted in some positive signals, this study nevertheless allowed clear separation between loci contributing selection and those that contributed to laboratory adaptation. Selection responses could then be clearly characterized in recently established lines from nature without results being driven by laboratory adaptation processes, which can be of concern (Harshman and Hoffmann 2000).

*Effective population size*

Effective population size (*Ne*) remained moderate across neutral markers for all replicate lines (Figure 3), indicating the study design successfully avoided extreme bottlenecks. Candidate SNPs exhibited strongly reduced *Ne*, as expected, with similar reductions shown at desiccation-candidate loci in desiccation-selected lines, and in lab-adaptation candidate loci in control lines. These patterns reflect the fact that selection has pushed changes in alleles beyond those expected under neutral expectations. Given the relatively independent nature of selection responses in the replicate lines, selection is therefore expected to drive divergence among replicate lines at the genomic level and despite the fact that the lines had similar phenotypes as a consequence of selection. Lines in each treatment also showed some *Ne* reduction at candidate loci for the other treatment. This makes sense for the desiccation-selected lines, which were also exposed to lab conditions, and although this probably did not exert selection as strong as the desiccation stress, it clearly also impacted changes in allele frequencies. However, it is not entirely clear why the lab-adaptation lines should show a reduction in *Ne* at the desiccation-selected candidate loci, unless there are pleiotropic effects between these traits.

*Gene length bias*

As expected, a positive relationship was detected between gene length and probability of sampling a gene in the simulation tests (Figure S22). The candidate gene lists generally tracked expectations but in some cases appeared skewed towards genes that were more likely to be detected in the simulations (e.g. Figure S22, panel E). This skew did not appear to be as severe in the desiccation-replicate lines, especially once the putative lab-adaptation genes were removed (panels O—V). However, many of the genes in the observed candidate gene lists for replicates D3 and D4 (those with the most SNPs) were also selected in over half the simulations. Further to this, the overlapping genes – those observed to occur in multiple replicate lists – appeared skewed towards genes that were more likely to appear in the simulated overlap lists (panels V—AW). Thus there are likely to be numerous false positives remaining in the data despite efforts to reduce these.

*Impact of coverage depth on SNP candidate status*

In all cases the candidate SNP coverage was not significantly lower and the was in fact in most cases higher than the random SNP coverage distribution (Figure S3). This confirms the stringency of the candidate SNP identification approach: loci were only considered significantly differentiated if they had reasonably high coverage depth, imparting accurate allele frequency estimates. However, we cannot rule out a possible ‘deflation’ in the number of candidate SNPs identified as shared among replicates, in areas that had low coverage in some replicates. This may explain why allele frequency change direction was almost always significantly shared across replicates (Figure 2), although sets of SNPs reaching ‘candidate’ status were not always significantly overlapping (Figure 1).

*Impact of multiple-mapping SNPs on gene-level overlap among replicates*

In this analysis we plotted the overlap (counts of shared candidate genes) among replicate combinations when considering only single-mapped genes, versus the overlap when considering only multi-mapped genes.

For the control replicates, we observed a slope much less than 1, indicating the overall patterns of gene overlap were strongly influenced by SNPs that mapped to multiple genes (Figure S18A). The vast majority of these SNPs were mitochondrial and this result is discussed elsewhere in the main text.

For the desiccation replicates, plotting the single-mapped gene overlap versus the multi-mapped gene overlap revealed no major effect of multi-mapping genes (Figure S18B). The slope of this relationship was close to 1, indicating gene overlap estimates were of a similar relative magnitude if considering only single-mapped genes, only multi-mapped genes, or all candidate genes together.

**Discussion**

*Gene length bias in shared candidate genes*

The 13 genes identified as candidates in all five desiccation replicates were *uif*, *Snoo*, *PAPLA*1, *Cdk*1, *aret*, *CG42313*, *CG10333*, *Atac*2, *MESR*3, *psq*, *CG9990*, *CG32373*, and *Bx*. The first nine of these occur on chromosome 2L, and *CG10333*, *Atac*2 and *MESR*3 are located in a contiguous stretch from 2L:18609576—2L:18612249. Many of these genes are very long, with all genes over 20 Kb in length except *CG10333* (2817 bp), *Atac*2 (2673 bp) and *Cdk*1 (1523 bp). Two genes were over 100 Kb in length (*aret*, 138 Kb; *CG32373*, 109 Kb). For this reason, and because this overlap list was no longer than expected by chance, we believe it likely includes false positive hits, with candidate SNPs repeatedly mapping to these genes due to their length (Figure S22). The enrichment in the *CG10333*—*Atac*2—*MESR*3 region is interesting, however. *CG10333* appears to be involved in regulation of RNA splicing; *Atac*2 is a component of the Ada2a-containing complex, which positively regulates transcription and facilitates action of some chromatin-remodelling complexes; and *MESR*3 appears to negatively regulate Ras signalling. Any (or all) of these genes could potentially affect downstream gene expression. *Cdk*1 also remains an intriguing candidate: it is relatively short, and yet appeared as a candidate gene in all five replicate lines. This gene encodes cyclin-dependent kinase 1, a crucial protein kinase controlling mitotic cell cycle stage transitions by phosphorylating numerous cellular proteins. Five genes were also predicted targets of candidate SNPs found to be eQTL/veQTLs in all replicates (Huang *et al.* 2015), *Acer*, *CG13743*, *CG9806*, *Dhc*16F and *Klp*98A. However, because these genes are all targeted by extremely high numbers of predicted eQTL/veQTLs across the genome (between 2974 and 9715 such SNPs each) and because the extent of gene-level overlap is no higher than expected by chance, we do not consider these five genes to be especially meaningful candidates.

**References**

Baldwin-Brown J. G., Long A. D., Thornton K. R., 2014 The power to detect quantitative trait loci using resequenced, experimentally evolved populations of diploid, sexual organisms. Mol. Biol. Evol. **31**: 1040–1055.

Bastide H., Betancourt A., Nolte V., Tobler R., Stöbe P., Futschik A., Schlötterer C., 2013 A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. PLoS Genet **9**: e1003534.

Bolker B., R Development Core Team, 2014 bbmle: Tools for general maximum likelihood estimation.

Burke M. K., Liti G., Long A. D., 2014 Standing genetic variation drives repeatable experimental evolution in outcrossing populations of *Saccharomyces cerevisiae*. Mol. Biol. Evol. **31**: 3228–3239.

Fabian D. K., Kapun M., Nolte V., Kofler R., Schmidt P. S., Schlötterer C., Flatt T., 2012 Genome-wide patterns of latitudinal differentiation among populations of Drosophila melanogasterfrom North America. Mol Ecol **21**: 4748–4769.

Fischer M. C., Rellstab C., Tedder A., Zoller S., Gugerli F., Shimizu K. K., Holderegger R., Widmer A., 2013 Population genomic footprints of selection and associations with climate in natural populations of Arabidopsis hallerifrom the Alps. Mol Ecol **22**: 5594–5607.

Franssen S. U., Nolte V., Tobler R., Schlötterer C., 2014 Patterns of linkage disequilibrium and long range hitchhiking in evolving experimental *Drosophila melanogaster* populations. Mol. Biol. Evol. **32**: 495–509.

Harshman L. G., Hoffmann A. A., 2000 Laboratory selection experiments using *Drosophila*: what do they really tell us? Trends Ecol Evol **15**: 32–35.

Huang W., Carbone M. A., Magwire M. M., Peiffer J. A., Lyman R. F., Stone E. A., Anholt R. R. H., Mackay T. F. C., 2015 Genetic basis of transcriptome diversity in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U.S.A.

Kapun M., van Schalkwyk H., McAllister B., Flatt T., Schlötterer C., 2014 Inference of chromosomal inversion dynamics from Pool-Seq data in natural and laboratory populations of *Drosophila melanogaster*. Mol Ecol **23**: 1813–1827.

Kessner D., Novembre J., 2015 Power analysis of artificial selection experiments using efficient whole genome simulation of quantitative traits. Genetics **199**: 991–1005.

Lynch M., Bost D., Wilson S., Maruki T., Harrison S., 2014 Population-genetic inference from Pooled-Sequencing data. Genome Biol Evol **6**: 1210–1218.

Tobler R., Franssen S. U., Kofler R., Orozco-terWengl P., Nolte V., Hermisson J., Schlötterer C., 2013 Massive habitat-specific genomic response in *D. melanogaster* populations during experimental evolution in hot and cold environments. Mol. Biol. Evol. **31**: 364–375.