Supporting Information

Methods

D. simulans experimental population and the selection regime

Ten replicates of *D. simulans* were established using 202 isofemale lines from a natural *D. simulans* population collected in Tallahassee, Florida, USA (*1*). These replicates were maintained in a new hot environment in which both temperature and light cycled every 12 hours between 18 and 28°C, corresponding to night and day. The replicates had a census population size of 1000 and ~50:50 sex ratio. The flies in each replicate were equally distributed across five 300ml bottles containing 70 ml of standard *Drosophila* medium.

Genome sequencing and mapping of sequence reads

Genomic DNA was extracted for all replicates at generation 0 (females only) and all evolved replicates in 10 generation intervals until generation 60 (mixed sexes). The replicates at generation 0 will be referred to as 'ancestral population' hereafter. Details of DNA extraction and library preparation will be provided upon request. Sequencing of paired-end 100bp reads resulted in an average genome-wide sequence coverage of ~216x for each ancestral and ~103x for each evolved replicate. Trimming, mapping and filtering of reads was performed as described in (*1*).

SNP calling

SNPs were called from replicates of the ancestral population; in brief, SNPs with base quality of 40 in at least one of the ten ancestral replicates were selected for further analyses. To improve the reliability of the pipeline, the polymorphic sites in the upper 1% and lower 1% tails of the coverage distribution (i.e.: ≥423x and ≤30x, respectively; upper tail based on the library with the highest sequencing depth, lower tail estimated from total coverage of all ancestral and evolved replicates at generation 60, Table S6) and minor alleles with coverage less than 10 reads were removed. Furthermore, we masked repeats (transposable elements were annotated using the pipeline described in (*2*)) and 5-bp regions flanking indels (identified by PoPoolation2 (*3*): using function *identify-genomic-indel-regions.pl* with options --*indel-window 5 --min-count 167*). The minimum read count cutoff corresponds to 2% of the average coverage across all ancestral and evolved replicates. We further masked 200-bp flanking the SNPs specific to autosomal genes translocated to the Y chromosome (*4*). The remaining 5,096,200 SNPs on the major chromosomes were used for subsequent analyses. For these SNPs we determined the allele frequencies using only reads with a base quality score of at least 20 at the SNP position.

Inference of candidate SNPs

To identify SNPs with pronounced allele frequency changes (AFC) Fisher's exact test and Cochran-Mantel-Haenszel (CMH) test were used. First, we contrasted the ancestral and the evolved replicates at generation 60 using the CMH test to identify SNPs with a consistent frequency change across replicates (using PoPoolation2, function *cmh-test.pl*). Second, the pronounced AFCs specific to each replicate were determined by contrasting each ancestral replicate with the corresponding evolved replicate at generation 60 (for example ancestral replicate 5 with evolved replicate 5) using Fisher's exact test (Popoolation2, function *fisher-test.pl*: option --*min-count 5*). For the parameters of *fisher-test.pl*, the minimum SNP coverage for each replicate was set to 5% of the average coverage of ancestral and evolved replicates at generation 60. For the maximum coverage, the upper 1% tail of the coverage distribution in the replicate with the highest sequencing depth was chosen (Table S6). In total, 10 Fisher exact tests were performed.

Neither the CMH nor the Fisher's exact test account for drift. Thus, to determine the candidate SNPs whose AFCs were higher than expected under neutral drift, forward Wright-Fisher simulations were performed with Nest (*5*) (function *wf.traj*) assuming independence among alleles. As a first step for the neutral simulations, the effective population sizes (N_e) were estimated for each of the replicates. We used windows of 1000 SNPs based on AFCs between the ancestral and evolved replicates at generation 60 for autosomes (each chromosome separately) and the X-chromosome using Nest (function *estimateWndNe*, method *Np.planI*). We averaged the medians of the N_e estimates across replicates; the estimated N_e was 291 for autosomes and 262 for the X-chromosome (Table S5) and we used these estimates of N_e to perform simulations to determine the False Discovery Rate (FDR: true positive/[false positive+true positive]) of the CMH test. We determined the FDR of Fisher's exact test by performing neutral simulation using replicate-specific N_e estimates for autosomes (Table S5) and a N_e of 262 for the X-chromosome. The simulation parameters (i.e. number of SNPs, allele frequencies in the ancestral replicates, sequence coverage of replicates, and the number of replicates and generations) matched the experimental data. Candidate SNPs were inferred based on an empirical CMH/Fisher's exact test cutoff using a 5% FDR based on neutral simulations.

Reconstruction of the selected haplotype blocks (selected alleles)

In experimental evolution studies using *D. melanogaster* many neutral SNPs are linked to the target(s) of selection (*6*, *7*). Therefore, it was proposed to shift the focus from candidate SNPs to haplotypes, which carry the selected target(s) (*8*). Taking advantage of correlated allele frequency changes of SNPs, which are characteristic for a given haplotype, the genomic region under selection can be inferred and the rising haplotype blocks reconstructed. We identified haplotype blocks using a modification of a recently published approach (*8*).

First, we grouped candidate SNPs together with stringent clustering (minimum average Pearson's correlation coefficient of 0.75 among SNPs). Candidate SNPs (5% FDR) inferred from both CMH and Fisher's exact tests were combined and used for haplotype block reconstruction; the candidate SNPs (5% FDR) from the CMH test were polarized based on the rising allele and only SNPs with a frequency increase of ≥0.2 (between any time points from F0 to F60) in at least two replicates were retained. Additionally, all candidate SNPs (5% FDR) from Fisher's exact tests were polarized based on the rising allele and used for block reconstruction. The allele frequencies of these candidate SNPs were transformed (arcsine of the square root, using *numpy.arcsin* and *numpy.sqrt* functions in Python) and standardized, i.e. centered to mean and scaled to unit variance (using function *sklearn.preprocessing.scale* in Python). The left and right arms of chromosomes 2 and 3 were concatenated for this analysis since some haplotype blocks overlap the centromere. All pairwise Pearson's correlation coefficients were computed among SNPs in sliding windows of 1Mb with a step size of 500kb. Each window should have at least a minimum of 20 SNPs. The correlation matrix was converted to Euclidean distance, d = √2(1-r), and entered in a distance matrix (using function *scipy.spatial.distance.squareform* in Python) which was used for hierarchical clustering (using function *scipy.cluster.hierarchy.linkage* in Python). SNPs with the minimum average Pearson's correlation coefficient of 0.75 were classified into a cluster. Only clusters with more than 20 SNPs were retained. Clusters obtained from overlapping windows were merged if they shared at least five SNPs.

Second, all the SNPs in clusters with average correlation of 0.75 (see above) were collected and used for another round of clustering (similar to above) but SNPs with the minimum average Pearson's correlation coefficient of 0.35 were classified into a cluster. Similar to above, clusters that share at least five SNPs in overlapping windows were merged.

Inference of selection coefficient

Selection coefficients (*s*) were estimated with Pool-Seq (*9*) (using the function *estimateSH*) which uses time-series allele frequency data to infer *s*. Because the selected SNP is not known, our estimates of *s* rely on the frequency of the selected haplotype block/allele (SI Text: Definition of selected allele). We defined the frequency trajectory of the selected haplotype block/allele in each replicate as the median frequency of all SNPs characteristic to the selected haplotype block/allele. A haplotype block/allele with a frequency increase of ≥0.1 is considered to be selected in a given replicate. An AFC of 0.1 corresponds to the upper 1% tail of the absolute average AFC distribution across 10 replicates in neutral simulations (see 'Inference of candidate SNPs' for details of neutral simulations). Thus, for replicates with at least 0.1 frequency increase of selected haplotype block/alleles after 60 generations *s* was estimated for each replicate separately, and the median *s* across the corresponding replicates is reported as the selected haplotype block/allele's *s* (Fig. 3B, Fig. S4A). This estimated *s* was used for subsequent analyses and simulations unless mentioned otherwise. To make sure that the estimated *s* is not biased by the method used for allele frequency estimation, we used different approaches to estimate the frequency of a given selected allele (SI Text: 'Different approaches to estimate allele frequencies for inference of selection coefficient'). A dominance of 0.5 was considered in all *s* estimations.

To identify factors contributing to the estimated *s,* we initially fitted a linear model with three fixed continuous effects (starting frequency: p_0, the number of replicates in which a specific allele increases ≥0.1 in frequency: replicate frequency*,* and locus size (the genomic region corresponding to an allele): *size*) and interaction between replicate frequency and p_0. *s* and replicate frequency were log10-transformed (*log10* function in R) and the square root of p_0 was arcsin-transformed (*asin and sqrt* functions in R). Due to non-linearity of p_0 a quadratic term (squared p_0) was also added to the model. The interaction between replicate frequency and p_0 was not significant and therefore dropped from the model ($s_{\mathrm{ijklm}} \sim \mu + p_{0\mathrm{i}}$ + replicate $\mathrm{frequency}_{\mathrm{j}} + {p_0}^2{}_{\mathrm{k}} + \mathrm{size}_{\mathrm{l}} + \mathrm{error}_{\mathrm{ijklm}}$). The data met the assumptions of normality of the residuals and homogeneity of variance.

Experimental inference of haplotypes

We experimentally determined the full chromosomal haplotypes by crossing males from the ancestral population and the evolved replicates with virgin females from the reference strain M252 (*10*) (Genbank BioSample SAMN02713493). In total, we obtained 100 haplotypes from five evolved replicates (20 from replicate 1 at F88, 36 from replicate 3 at F103, 20 from replicate 4 at F88, 12 from replicate 7 at F103, 12 from replicate 10 at F103). We also obtained

189 haplotypes from isofemale lines that were used to seed the ancestral population. A single F1 female of each cross was used for DNA extraction and sequencing. Details of DNA extraction and library preparation are available upon request. Trimming the sequencing reads for evolved and ancestral haplotypes was performed as described in (*1*). Reads of the evolved haplotypes were mapped as explained in (*1*) and for the ancestral haplotypes using Novoalign version 3.03.2 (http://www.novocraft.com/products/novoalign/, parameters: -r Random –i mean,standard deviation of reads). Filtering of mapped reads for evolved and ancestral haplotypes was performed as described in (*1*).

Evolved haplotypes were called from F1 individuals as described in (*11*). The F1 genotype was compared to the reference strain (Genbank assembly accession: GCA_000820565.1) and the parental alleles were determined. Haplotype base calls were performed for positions that had a coverage larger than 10 and less than the maximum 2% coverage of the respective library. For ancestral haplotypes, SNP were called with freebayes (*12*) (version v1.1.0-46-g8d2b3a0) and the recombination rate was computed using LDJump (*13*) (version 0.1.4).

Contrasting selective sweep model and quantitative genetics model

We observed a highly heterogeneous response across the 10 hot-evolved replicates; most of the 99 selected alleles increase in only four to six replicates (Fig. 3C,D, Fig. 4B,D). We used this heterogeneous pattern among replicates to discern several different adaptive scenarios (Fig. S3).

To compare different models (sections A-D below correspond to scenarios A-D in Fig. S3) to the empirical data, we used the 'replicate frequency spectrum' (RFS), i.e. the frequency distribution of replicates in which selected alleles increase in frequency, as a summary statistic to measure the fit between simulated and observed data. Specifically, if an allele increases in frequency by at least 0.1 after 60 generations, we considered this a replicate with selection signature (an AFC of 0.1 corresponds to the upper 1% tail of the absolute average AFC distribution across 10 replicates in neutral simulations). In all simulations summarized in section A-D, RFS was determined, and for each category, mean and 95% confidence interval were computed. For each simulation, we compared RFS in observed data and the simulations by a χ^2 goodness of fit test with n-2 degree of freedom (n=number of replicates).

A. Sweep model without linkage and a constant selection coefficient across replicates

Because drift in small populations can result in considerable heterogeneity among replicates, we used computer simulations to test whether the observed heterogeneity among replicates can be explained by the combined effect of the sweep paradigm and drift. A total of 1000 sets of forward Wright-Fisher simulations were performed using Nest (function *wf.traj*). In each set, 99 independent alleles (matching the observed number of selected alleles) were simulated using the observed starting frequency (median frequency of SNPs characteristic to the selected haplotype block/allele) in the ancestral replicates (Fig. 3A), replicate-specific N_e (Table S5) and allele-specific s (Fig. 3B, Fig. S4A) assuming no linkage and epistasis (A2 in Fig. S3). Selected alleles on autosomes and the X-chromosome were simulated separately using N_e estimated for autosomes and the X-chromosome, respectively (Table S5). The allele-specific s was estimated by the default approach described in Materials and Methods: Inference of selection coefficient; in brief, for each selected haplotype block/allele, we first estimated the median frequency trajectory across all characteristic SNPs in a given replicate. Replicates with AFC ≥0.1 were used to determine *s,* and the median *s* across replicates was used in the simulations. Simulations with slightly different parameters were performed to assure that the above simulations are not biased by the allele-specific starting frequency and *s* (SI Text, Fig. S3, S5).

B. Sweep model with linkage and a constant selection coefficient across replicates

We simulated a sweep model assuming linkage among selected alleles to rule out that Hill-Robertson interference caused the genetic heterogeneity among the replicates. We used 189 individually sequenced haplotypes from the ancestral population (Materials and Methods: Experimental inference of haplotypes) for the simulations. We simulated 99 selected alleles (the number of selected alleles in the experimental data) in 10 replicates of a population of 300 diploids (corresponding to the estimated N_e) in 1000 iterations with the recombination rate estimated from the haplotypes. The allele-specific *s* were obtained from the median *s* (Fig. S4A) estimated from replicates with AFC ≥0.1 for the selected alleles (B2 in Fig. S3). For the simulations, the position and starting frequency of the selected target was chosen to match the position of one of the characteristic SNPs in the selected haplotype block/allele and the starting frequency of the selected allele in the ancestral population (Fig. 3A), respectively. Simulations were performed using function *w* in Mimicree2 (version mim2-v193), which uses haplotype information and also accounts for the differences between the X-chromosome and autosomes (https://sourceforge.net/projects/mimicree2/).

Simulations with different parameters were performed to assure that the above simulations are not biased by the allele-specific starting frequency and *s*. We simulated alleles using *s* (Fig. S4F) and starting frequency (Fig. S4H) estimated for the core region of alleles with ≥0.1 frequency increase after 60 generations (B1 in Fig. S3, see SI Text: 'Different approaches to estimate allele frequencies for inference of selection coefficient' for definition of core region). All other parameters are similar to above simulations. We did not obtain a good fit to the empirical Replicate Frequency Spectrum (Fig. S8).

C. Genetic redundancy model

One prediction of the quantitative trait model with stabilizing selection is that the same trait optimum can be obtained by different combinations of contributing alleles (genetic redundancy). Rather than simulating the trajectories of selected alleles we first tested whether we find evidence for genetic redundancy.

Independent of whether we conditioned on a 0.1 or 0.2 allele frequency increase after 60 generations, a different number of selected alleles was detected among the replicate populations (Fig. S9). Nevertheless, all replicates converged for the high-level phenotypes: fitness, fat content and resting metabolic rate (Materials and Methods: Phenotypic assays; Fig. 2C,D, Fig. S1B). Therefore, we tested if the frequency distribution of the selected alleles among replicates of our experiment fits a model of full genetic redundancy. Assuming that all alleles are functionally equivalent, we generated 1000 data sets using delete-*d* jackknifing. In each set the number of selected alleles for each replicate matched our observations (Fig. 4A), but was randomly drawn (without replacement) from the total of 99 selected loci (C in Fig. S3).

We calculated the Jaccard index ($|A \cap B| / |A \cup B|$) of all pairwise combinations of hot-evolved replicates using the number of selected alleles with ≥0.1 AFC after 60 generations to determine the similarity among replicates (C in Fig. S3). Correspondingly, pairwise Jaccard indices were also computed for simulations of genetic redundancy model (Fig. 4B).

As a further test for the redundancy model, we predicted the fitness for all 10 hot-evolved replicates by summing frequencies of all the 99 selected alleles at generation 60 (C in Fig. S3). Because this test assumes equal effects of the selected alleles, similar values for the predicted fitness across replicates supports a redundancy model with alleles of similar effects.

D. Quantitative trait model

To determine whether the genomic heterogeneity among the replicates of the empirical data could be obtained by a quantitative trait model, we simulated frequency trajectories of alleles contributing to a quantitative trait after a change in trait optimum (D in Fig. S3). Using forward simulations in diploids assuming random mating, we simulated 1000 iterations of a quantitative trait in 10 replicates with 99 contributing alleles having the same starting frequency as the selected alleles in the empirical data (Fig. 3A). Alleles were in linkage equilibrium and had equal effects. The fitness ranged between 0.5-4.5 and the mean fitness optimum was set to 0.6±0.3 (standard deviation). Simulations were performed using the python script (*frequencyAt-pheno-quantitative.py*) provided in (*14*).

Identification of transposable elements

Because the ancestral population experienced the invasion of a P-element (*15*), it may be possible that new P-element insertions were driving adaptation and thus contribute to the observed heterogeneity among replicates. Adaptation driven by the P-element should result in a frequency increase of the P-element, which matches the frequency increase of the corresponding allele in a given replicate. Hence, we first identified P-element insertions, and then compared the frequency increases of the P-elements to the one of selected alleles.

The raw reads of all 10 evolved replicates at generation 60 were separately mapped to a TE-merged-reference genome using bwa (*16*) version 0.7.9a (*bwasw* algorithm). The reference genome consists of the repeat-masked reference genome and TE sequences as described in (*2*). The paired reads were restored (function *se2pe*) and a ppileup file was generated using PoPoolationTE2 (*17*). All 10 evolved replicates were down-sampled (function *subsamplePpileup*) and transposable element (TE) insertions were identified with PoPoolationTE2 (functions *identifySignatures* and *frequency*). The identified TE insertions were filtered (*filterSignatures* with parameters -*-max-otherte-count 2 --max-structvar-count 2*) and the final set of insertions was identified (function *pairupSignatures*).

We filtered the identified P-elements by selecting those with frequencies ≥0.1 that have insertion sites in the genomic regions corresponding to selected alleles. Furthermore, the P-elements with frequency increase must be present in at least one of the replicates which have frequency increase of ≥0.1 for any given selected allele. Then, for each selected allele with an available P-element in the filtered dataset we computed delta (absolute frequency difference) using the frequency of the allele in any of the two replicates with highest frequency at generation 60 and the frequency of the P-element in those replicates.

Gene Ontology (GO) and pathway enrichment analyses

We used Gowinda (*18*) to associate the characteristic SNPs in the selected haplotype blocks/alleles with their biological functions and determine enrichment of any specific GO category. This tool corrects for biases introduced by gene lengths. The GO enrichment analysis was performed in gene mode, with 100,000 simulations. The associated GO terms were downloaded from GoMiner (*19*). We repeated the enrichment analysis using pathways obtained from KEGG (*20*).

Phenotypic assays

All phenotypic assays were performed in a common garden setting (with temperature fluctuating between 18 and 28°C) to eliminate the possibility that uncontrolled environmental variation is affecting the phenotypic measurements. We reconstituted the ancestral population from the isofemale lines that were used to seed the original ancestral population. As the effective population size of isofemale lines is very small, very limited adaptation is expected to have occurred during their maintenance in the lab (*21*, *22*). Even mutations accumulated during the maintenance of the isofemale lines are not expected to have a major influence as they will be specific to individual isofemale lines and will thus represent only a small fraction in the reconstituted ancestral population. We refer to the reconstituted ancestral population as 'ancestral population' hereafter. All 10 evolved replicates and the ancestral population were maintained for at least two generations at the assaying conditions with controlled density (400 eggs per bottle) prior to assays to avoid maternal effects.

A. Fecundity assay

At generation 103, between three to six technical replicates were set up for each of the 10 evolved replicates to ensure reliable fecundity estimates. 10 technical replicates were set up for the ancestral population. Immediately after eclosion, around 250 flies (females and males) were put in a bottle. Since the flies were collected under CO_2 anesthesia, we only started measuring fecundity in two days old flies. For the next four days, which corresponds to the egg laying period during the experimental evolution, the flies were daily transferred to new bottles without CO_2 anesthesia and the eggs in each bottle were counted. After the fourth day, females and males were separated, the females were counted, dried and weighed. Fecundity, measured as the total number of eggs laid per female during four days (between day two to five after eclosion), was log10-transformed and analyzed using linear model.

To test for the significance between fecundity of the ancestral and evolved (all 10 evolved replicates combined) populations, we initially fitted a linear mixed model with a fixed categorical effect (population) with two levels (ancestral and evolved) and a fixed continuous effect (average body weight). The technical replicates were included as random effect. The mixed model was not significantly different from the linear fixed effects model tested using the function *anova*. Therefore, following the principle of parsimony we present results from the linear fixed effects model ($\text{Fecundity}_{ijk} \sim \mu + \text{population}_i + \text{average body weight}_j + \text{error}_{ijk}$). The data met the assumptions of normality of the residuals and homogeneity of variance.

The model to test for differences between the 10 evolved replicates was the same as above, but with population being a fixed categorical effect with 10 levels. The mixed effects model including a random effect to model the covariance between the technical replicates was not significantly better than the fixed effects model and hence we dropped it in favor of the simpler fixed effects model. Significance of the fixed effects was tested using ANOVA F-tests. We present effect sizes as lsmeans calculated with package *lsmeans* (*23*) and used Tukey's honest significant difference (HSD) to correct for multiple testing.

B. Resting metabolism assay

The evolved populations were at generation 113 when phenotyped for resting metabolic rate. The flies used for measurement of metabolic rate were maintained at controlled density as described above (evolved replicates for seven to nine and ancestral population for four to seven generations). Flies were collected immediately after eclosion, mated for 24 hours and then females and males were separated and maintained in bottles with *Drosophila* medium (150 flies in a bottle). After 48 hours of CO_2 anesthesia recovery, flies were used for metabolic rate measurement. Females and males were four to five and six to seven days old, respectively, during resting metabolic measurements. Our preliminary assays identified no significant difference in the metabolic rate of four to five and six to seven days old males when groups of 25 males were tested in RC respirometry chambers (30ml, Sable Systems, Las Vegas, Nevada, USA). For measuring metabolic rate, 150 flies were transferred to 250-ml bottles without CO_2 anesthesia. To avoid desiccation and starvation, 50 ml of Drosophila medium were placed in each bottle and sealed with a stopper. The resting metabolic rate was measured by repeated

CO_2 emission measurements of stop-flow respirometry (Sable Systems, Las Vegas, Nevada, USA). Flies of different replicates were randomly assigned to each bottle and one bottle with only *Drosophila* medium was used as an empty control in each run. CO_2 measurement was conducted at 18°C in the dark, overnight for at least 12hrs. During the measurement assay, an 8-channel-multiplexer (RM8 Intelligent Multiplexer) controlled the sequential flushing and closing of eight bottles. Each bottle was flushed for 15 minutes at a constant flow rate of 75 µL/min. After the flush phase, the bottle was closed while the rest of bottles continued to be flushed. Therefore, CO_2 in each bottle was measured every two hours. The flushed air passed through a Magnesium perchlorate column to remove water, and CO_2 was measured with a CA-10A Carbon dioxide Analyzer (Sable Systems, Las Vegas, Nevada, USA). At the end of each run, the flies were counted, dried and weighed. An in-house macro was used for computing the total CO_2 emission during each flushing time and the mean flow rate by ExpeData software. The resting metabolic rate for each bottle was computed as the average of the three lowest data points (*24*). The metabolic rate is presented as V_{CO2} $\mu L\ h^{-1}\ mg^{-1}$. A total of 32 measurements (16 for each sex) were done for the ancestral population. At least six measurements were performed for each evolved replicate (three for each sex).

To test for the significance between the metabolic rate of the ancestral and evolved (all 10 evolved replicates combined) populations, we initially fitted a linear mixed model with two fixed categorical effects (population and sex) each with two levels (population: ancestral and evolved, sex: female and male) and interaction between the fixed categorical effects. Multiple measurements of replicates were included as random effect. Similar to fecundity, the mixed model was not significantly different from the linear fixed effects model. Therefore, we present results from the linear fixed effects model (Metabolism$_{ijk}$ ~ μ + population$_i$ + sex$_j$ + population$_i$: sex$_j$ + error$_{ijk}$). The data met the assumptions of normality of the residuals and homogeneity of variance.

We used the same model as above to test for differences between the 10 evolved replicates, but with population being a fixed categorical effect with 10 levels. We included a random effect to model covariance between the technical replicates but it was not significantly better than the fixed effects model and hence dropped in favor of the simpler fixed effects model. Significance of the fixed effects was tested using ANOVA F-tests. We present effect sizes as lsmeans and used Tukey's HSD to correct for multiple testing.

C. Body fat assay

We assayed the ancestral population and evolved replicates for fat content at generation 124. Similar to other phenotypic assays, the evolved replicates and ancestral population were maintained in a density controlled common garden setting (for three to six generations). After eclosion flies were mated, females and males were separated, and placed in vials (eight flies in each vial). After 48 hours, the body fat was measured in four-day-old flies three hours after the start of the daily 28ºC cycle. Homogenates were prepared as described in (*25*) and lipid measurements were performed by coupled colorimetric assay as described in (*26*). Flies were placed in a 2-ml screwcap tube containing 600 μl 0.05% Tween-20 and two 5-mm sterile metal beads and homogenized by a SPEX SamplePrep 1600 MiniG for 2 minutes at 1,475 rpm. Homogenates were heat-inactivated for 5 minutes at 70°C, centrifuged for 9 minutes at 9,000 rpm, and 200 μl of the supernatant was transferred to an Eppendorf tube and immediately used for lipid measurements. We used a Glycerol standard (Sigma G7793) as reference. 50μl of supernatant and standards (1.2, 1, 0.8, 0.6, 0.4, 0.2, 0.1, 0 mg/ml) were transferred to a 96-well plate and blank absorbance was measured at 540 nm in an EnSpire 2300 Microplate Reader (PerkinElmer). 200 μl of Triglyceride Working Reagent (Sigma, Catalog number TR0100) was added to each sample and standards and incubated at 37°C for 30 minutes with mild shaking. After the incubation time, the final absorbance was measured at 540 nm. The two measurements were first blank-corrected by subtraction of blank (i.e. 0.05% Tween-20) and then the second absorbance was subtracted from the first absorbance. Glycerol standards were done in duplicates and two measurements were averaged for the standard curves with polynomial regression line to compute the concentration of triglyceride (TAG) in the samples. Fat content is expressed as μg TAG equivalents/fly. A total of 26 measurements (13 for each sex) were done for the ancestral population and eight measurements were performed for each evolved replicate (four for each sex).

The measured fat content (μg TAG equivalents/fly) was log-transformed and analyzed using a linear model. To test for the significance between the fat content of the ancestral and evolved (all 10 evolved replicates combined) populations, we initially fitted a linear mixed model with two fixed categorical effects (population and sex) each with two levels (population: ancestral and evolved, sex: female and male), a fixed continuous effect (average body weight) and interaction between the fixed categorical effects. Multiple measurements of replicates were included as random effect. Due to error during weighing, for six measurements we did not provide weight values in the linear model. The average body weight was not significant and therefore dropped from the model. The mixed model was also not significantly different from the linear fixed effects model. Therefore, we present results from the linear fixed effects model ($\text{Lipid}_{ijk} \sim \mu + \text{population}_i + \text{sex}_j + \text{population}_i : \text{sex}_j + \text{error}_{ijk}$). The data met the assumptions of normality of the residuals and homogeneity of variance.

The same model as above was used to test for differences between the 10 evolve replicates. Here, we treated population as a fixed categorical effect with 10 levels. Similar to other phenotypic assays, the mixed effects model including a random effect was not significantly better than the fixed effects model and hence dropped in favor of the simpler fixed effects model. The interaction between fixed effects was dropped from the final model since it was not significant. Significance of the fixed effects was tested using ANOVA F-tests. We present effect sizes as lsmeans and used Tukey’s HSD to correct for multiple testing.

All the analysis and data visualization have been performed in Python 2.7.10 (Python Software Foundation. Python Language Reference) and R 3.3.1 (R development Core Team. 2015).

SI Text

Reconstructing blocks of selected haplotypes

Because only few recombination events occur during 60 generations of E&R in *Drosophila*, selected variants typically occur on rather large haplotypes (*6*, *7*). While some methods have been proposed to reconstruct haplotypes from PoolSeq data (*27*–*29*), in absence of information about all ancestral haplotypes, the inference does not have sufficient reliability. Although we have sequenced 189 phased haplotypes from the ancestral population, this number accounts for only about 25% of the founder chromosomes (the 202 isofemale lines used for establishing the ancestral population were not inbred). Hence, we identify haplotypes carrying the beneficial mutation using a modification of a recently published approach (*8*).

We identified 5,096,200 SNPs on the major chromosomes after stringent filtering steps (Materials and Methods: SNP calling). To identify SNPs with pronounced AFC, we performed CMH and Fisher's exact tests contrasting the ancestral replicates with the evolved replicates at generation 60 (Materials and Methods: Inference of candidate SNPs). However, these test do not account for drift. Thus, to identify SNPs with more pronounced AFC in the evolved replicates than expected by genetic drift, we performed forward Wright-Fisher simulations. Based on an empirical CMH/Fisher's exact test cutoff using a 5% FDR we obtained 47,532 candidate SNPs across replicates (CMH test) and 4,667 additional SNPs deviated from neutral expectations in at least one of the replicates (Fisher's exact test). This number of candidate SNPs is heavily inflated due to the considerable linkage expected in our experiment.

We reasoned that SNPs specific to selected haplotypes have correlated allele frequencies across replicates and time points (*8*). Thus, we clustered SNPs by allele frequencies and shifted our focus from candidate SNPs to selected haplotypes. First, we grouped SNPs with a stringent correlation (average Pearson's correlation=0.75, Materials and Methods: Reconstruction of the haplotype blocks (selected alleles)) and identified many relatively short haplotype blocks (Fig. 1A, B). Recombination, either in the ancestral population or during the experiment results in less strongly correlated allele frequency trajectories. Hence, several adjacent haplotype blocks show very similar frequency change across replicates (Fig. 1C). We combined such short blocks with similar trajectories using less stringent clustering (average Pearson's correlation=0.35). The importance of combining SNPs with correlated allele frequency trajectories into a haplotype block is evident from the presence of several distinct peaks in the genomic region underlying a haplotype block (Fig. 1A) but a naïve interpretation may have suggested several independent alleles. A total of 99 selected haplotype blocks containing 23,835 SNPs were identified on the major chromosomes.

To demonstrate the robustness of this approach, we sequenced 100 phased haplotypes from five different evolved replicates and also 189 phased haplotypes from the ancestral population. 84% and 82% of the identified haplotype blocks were validated in the evolved and ancestral haplotypes, respectively (see Fig. 1F for an example, Table S1). In total 96% of the identified haplotype blocks were experimentally validated. Blocks that could not be validated either have low frequency in the ancestral population or had frequency increase in an evolved replicate with no available phased haplotype.

The reconstructed haplotype block is considered a selected allele for the subsequent analyses (Fig. 1D, E).

Definition of selected allele

Typically, multiple haplotypes are associated with a putative selection target, in particular when it occurs at higher frequency. As we cannot identify the selected mutation, we use the term “selected allele” to describe a suite of haplotype blocks carrying the target(s) of selection. Due to the limited mapping resolution in our experiment, we cannot distinguish whether the selection signature of a selected allele is generated by multiple selected targets (allelic heterogeneity) or a single target of selection.

Nevertheless, we addressed the possibility that the strong selection response in our experiment is driven by the combined effect of a large number of small effect alleles located on the selected haplotype blocks: if a large number of randomly distributed alleles contributes to the selected trait (i.e. infinitesimal model), the size of a selected genomic region should correlate with the number of contributing alleles-resulting in stronger selection for larger haplotype block. We tested this by regressing the estimated selection coefficient (*s*) with haplotype block’s starting frequency, replicate frequency (the number of replicates in which a specific haplotype block increases ≥0.1 in frequency), and locus size (the genomic region corresponding to a haplotype block). Because only the starting frequency ($p < 1.11\text{e-}12$) and replicate frequency ($p < 2.2\text{e-}16$) were significant, our data do not support a model with many randomly distributed targets of selection. The model explains 80% of the variance. Nevertheless, with non-random distribution of selection targets the results may differ.

Different approaches to estimate allele frequencies for inference of selection coefficient

Our estimates of selection coefficient (*s*) rely on the frequency of the selected allele because the causal SNP is not known. We defined the frequency of a selected haplotype block/allele as the median frequency of all SNPs characteristic to the haplotype block/allele. By default we estimated *s* for each allele as described in Materials and Methods: Inference of selection coefficient. Specifically, if a specific haplotype block/allele increases in frequency after 60 generations by at least 0.1 in a replicate we considered this as a replicate with selection signature. We used 0.1 AFC as a threshold because it corresponds to the upper 1% tail of the absolute average AFC distribution across 10 replicates in neutral simulations (Materials and Methods: Inference of candidate SNPs). However, to assure the estimated *s* is not biased by the method used for determining the frequency of the selected allele, different approaches were used to estimate the allele frequency.

First, we set a more stringent threshold for significant frequency increase after 60 generations: 0.2. Similar to the default method, *s* was estimated for all replicates with at least 0.2 frequency increase after 60 generations, and the median *s* across replicates was reported (Fig. S4C). Second, we reasoned that *s* might be underestimated because the selected haplotype block/allele includes a broad genomic region and SNPs with different starting frequencies (Fig. 1, Materials and Methods: Reconstruction of the haplotype blocks (selected alleles)). To avoid this bias we assumed that the region with the highest *s* in a haplotype block, i.e. allele, contains the target(s) of selection: each haplotype block/allele was divided into several regions (clustering with stringent correlation coefficient of 0.75) and for each region the median *s* across the replicates with a frequency increase ≥ 0.1 was computed, similar to the default method. The region with the largest *s* was chosen as the core region of the haplotype block/allele, and *s* (Fig. S4F) and the starting frequency (Fig. S4H) of the core region was used as the representative of the selected haplotype block/allele.

Furthermore, we checked whether restricting the calculation of the median *s* to those replicates with significant frequency increase biases the estimate of *s*. Rather than reporting the median *s*, we also reported *s* separately for each replicate. We show the replicate-specific *s* for all approaches of estimating *s* (Fig. S4B, D, G: 0.1 AFC, 0.2 AFC, core region of allele, respectively). Overall, all methods agree very well (Fig. S4), but an AFC $\geq$0.1 is the most conservative method resulting in the lowest *s*.

Simulation to test for a sweep model without linkage and a constant selection coefficient across replicates and time

Simulations of a sweep model assuming no linkage and epistasis with the default parameters (A2 in Fig. S3) specified in Materials and Methods (the observed starting frequency: Fig. 3A, replicate-specific N_e: Table S5 and allele-specific *s*: Fig. 3B) did not provide a good fit to the observed heterogeneity in the experimental data (Fig. 3C). To assure that the results of simulations are not biased by the allele-specific starting frequency and *s*, additional simulations were performed (Fig. S3).

Selection coefficient is a key parameter for our simulations, but it depends critically on the estimated frequencies. Given the uncertainty about the starting frequency of the selected allele, we tested the robustness of our conclusions by additional simulations using different estimates of *s* as detailed in: SI, Different approaches to estimate allele frequencies for inference of selection coefficient).

We performed simulations of a sweep model (A4 in Fig. S3) using *s* estimated with a more stringent threshold (0.2, Fig. S4C) for significant frequency increase after 60 generations (Fig. S5A). Furthermore, we simulated alleles using *s* (Fig. S4F) and starting frequency (Fig. S4H) estimated for the core region of alleles with ≥0.1 frequency increase after 60 generations (A1 in Fig. S3, Fig. S5B). Moreover, we simulated alleles using *s* and starting frequency estimates for the core region of alleles with ≥0.2 frequency increase (A3 in Fig. S3, Fig. S5C).

20 of the 99 observed selected alleles in the experimental evolution increased in frequency ≥0.1 in one to three replicates only. These alleles have low starting frequencies (Fig. S7A) and the highest estimated *s* (Fig. S7B). This may imply that we overestimated *s* in these replicates (winner's curse). To exclude the possibility that an overestimate of *s* for alleles which increased in frequency only in a small number of replicates affects the outcome of simulations, we repeated the simulations with the default parameters (the observed starting frequency: Fig. 3A, replicate-specific N_e: Table S5 and allele-specific *s*: Fig. 3B) using only those 79 alleles that increased ≥0.1 in frequency in more than four replicates (A5 in Fig. S3, Fig. S5D).

We estimated *s* using the frequency trajectory of replicates with ≥0.1 increase for the selected alleles. This might result in overestimation of *s* and bias the simulation results. To test whether using the frequency trajectory of all replicates for *s* estimation affects the simulation results, we performed simulations for these 79 alleles using *s* estimated from the median frequency change in all 10 replicates even if the frequency increase was less than 0.1 (A6 in Fig. S3, Fig. S5E).

Independent of the estimate of *s*, we could not obtain a good fit to the empirical replicate frequency spectrum (Fig. S5).

References

1. Barghi, N., Tobler, R., Nolte, V. & Schlötterer, C. *Drosophila simulans* : A species with improved resolution in Evolve and Resequence studies. *G3*, 7, 2337–2343 (2017).
2. Kofler, R., Nolte, V. & Schlötterer, C. Tempo and mode of transposable element activity in *Drosophila*. *Plos Genetics*, 11(7): e1005406 (2015).
3. Kofler, R., Pandey, R. V. & Schlötterer, C. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* 27, 3435–6 (2011).
4. Tobler, R., Nolte, V. & Schlötterer, C. High rate of translocation-based gene birth on the *Drosophila* Y chromosome. *Proc. Natl. Acad. Sci. U.S.A.*, 114 (44): 11721-11726 (2017). doi:10.1073/pnas.1706502114.
5. Jónás, Á., Taus, T., Kosiol, C., Schlötterer, C. & Futschik, A. Estimating the effective population size from temporal allele frequency changes in experimental evolution. *Genetics* 204, 723–735 (2016).
6. Tobler, R. *et al.* Massive habitat-specific genomic response in *D. melanogaster* populations during experimental evolution in hot and cold environments. *Mol. Biol. Evol.* 31, 364–75 (2014).
7. Franssen, S. U., Nolte, V., Tobler, R. & Schlötterer, C. Patterns of linkage disequilibrium and long range hitchhiking in evolving experimental *Drosophila melanogaster* populations. *Mol. Biol. Evol.* 32, 495–509 (2015).
8. Franssen, S. U., Barton, N. H. & Schlötterer, C. Reconstruction of haplotype-blocks selected during experimental evolution. *Mol. Biol. Evol.* 34, 174-18 (2016).
9. Taus, T., Futschik, A. & Schlötterer, C. Quantifying selection with Pool-Seq time series data. *Mol. Biol. Evol.* 34(11), 3023-3034 (2017).
10. Palmieri, N., Nolte, V., Chen, J. & Schlötterer, C. Genome assembly and annotation of a *Drosophila simulans* strain from Madagascar. *Mol. Ecol. Resour.* 15, 372–81 (2015).
11. Kapun, M., van Schalkwyk, H., McAllister, B., Flatt, T. & Schlötterer, C. Inference of chromosomal inversion dynamics from Pool-Seq data in natural and laboratory populations of *Drosophila melanogaster*. *Mol. Ecol.* 23, 1813–27 (2014).

12. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 (2012).

13. Hermann, P., Heissl, A., Tiemann-boege, I. & Futschik A. bioRxiv 190876; doi: https://doi.org/10.1101/190876 (2017)

14. Franssen, S. U., Kofler, R. & Schlötterer, C. Uncovering the genetic signature of quantitative trait evolution with replicated time series data. *Heredity*, 118(1), 42-51 (2017).

15. Kofler, R., Hill, T., Nolte, V., Betancourt, A. J. & Schlötterer, C. The recent invasion of natural Drosophila simulans populations by the P-element. *Proc. Natl. Acad. Sci. U.S.A.* 112(21), 6659-6663 (2015).

16. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–60 (2009).

17. Kofler, R., Gomez-Sanchez, D., & Schlötterer, C. PoPoolationTE2: Comparative population genomics of transposable elements using Pool-Seq. *Mol. Biol. Evol.* 33(10), 2759-64 (2016).

18. Kofler, R. & Schlötterer, C. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics* 28, 2084–5 (2012).

19. Zeeberg, B. R. *et al.* GoMiner : a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* 4:R28, (2003).

20. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–62 (2016).

21. Tobler, R., Hermisson, J. & Schlötterer, C. Parallel trait adaptation across opposing thermal environments in experimental *Drosophila melanogaster* populations. *Evolution* 69, 1745–59 (2015).

22. Nouhaud, P., Tobler, R., Nolte, V. & Schlötterer, C. Ancestral population reconstitution from isofemale lines as a tool for experimental evolution. *Ecol. Evol.* 6, 7169–7175 (2016).

23. Lenth, R. V. Least-Squares Means: The R package lsmeans. *J. Stat. Softw.* 69, (2016).

24. Jensen, P., Overgaard, J., Loeschcke, V., Fristrup, M. & Malte, H. Inbreeding effects on standard metabolic rate investigated at cold , benign and hot temperatures in *Drosophila melanogaster*. *J. Insect Physiol.* 62, 11–20 (2014).

25. Gáliková, M. *et al.* Energy homeostasis control in *Drosophila* adipokinetic hormone mutants. *Genetics* 201, 665–83 (2015).

26. Hildebrandt, A., Bickmeyer, I. & Kühnlein, R. P. Reliable *Drosophila* body fat quantification by a coupled colorimetric assay. *PLoS One* 6, e23796 (2011).

27. Kessner, D., Turner, T. L. & Novembre, J. Maximum likelihood estimation of frequencies of known haplotypes from pooled sequence data. *Mol. Biol. Evol.* 30, 1145–58 (2013).

28. Long, Q. *et al.* PoolHap: inferring haplotype frequencies from pooled samples by next generation sequencing. *PLoS One* 6, e15292 (2011).

29. Burke, M. K., Liti, G. & Long, A. D. Standing genetic variation drives repeatable experimental evolution in outcrossing populations of Saccharomyces cerevisiae. *Mol. Biol. Evol.* 31, 3228–39 (2014).

SI Figures

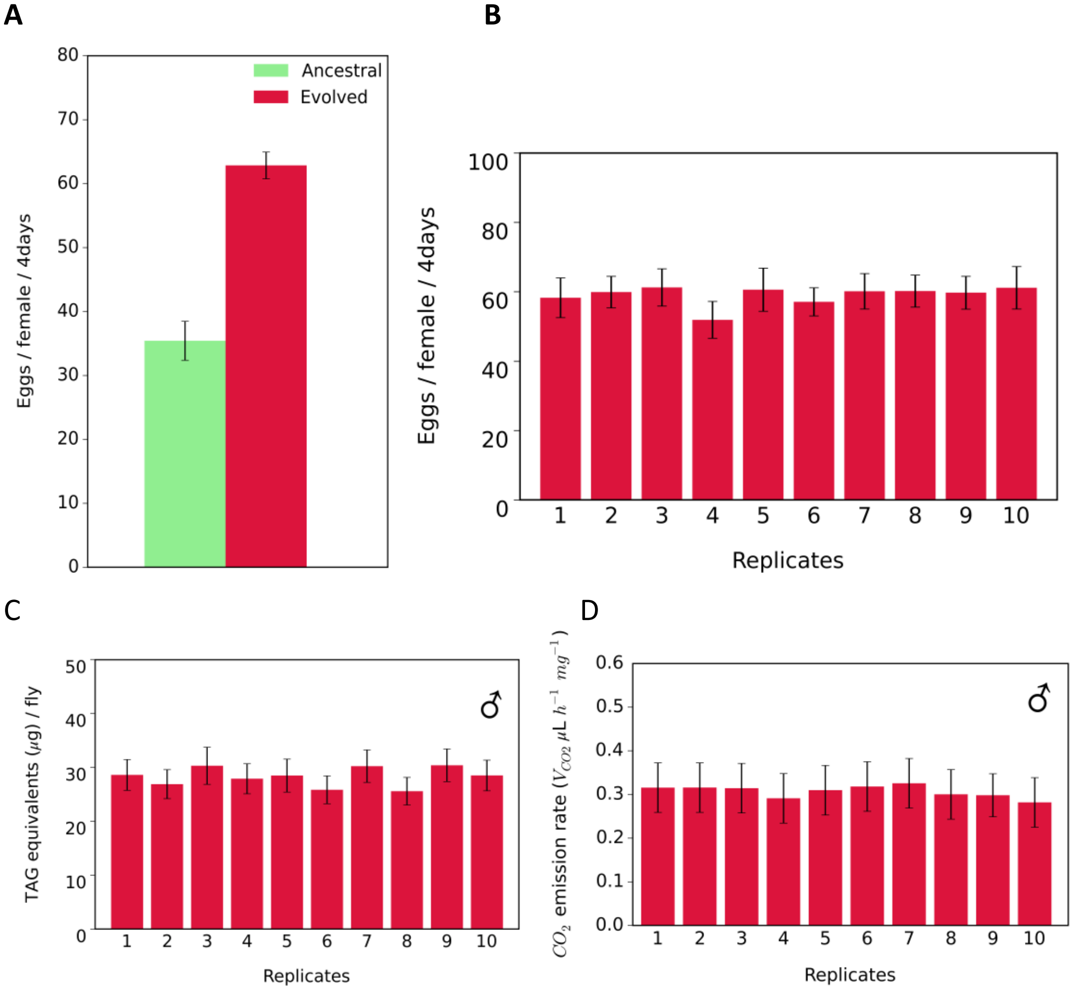

Fig. S1 Increased fitness and phenotypic similarity among 10 hot-evolved replicates. A) Hot-evolved females are more fecund than the ancestral population (ANCOVA, Tukey’s HSD test *p*-value < 0.0001). The number of eggs laid over four days (two to five days after eclosion) were counted B) Females of 10 hot-evolved replicates are equally fecund (ANCOVA, Tukey’s HSD test, *p*-value > 0.05). Similar fat content (C) and metabolic rate (D) were measured among males of the hot-evolved replicates (Two-way ANOVA, Tukey’s HSD test *P* > 0.05). The bars show least-squares means of the linear model and error bars depict 95% confidence levels of least-squares means (’lsmeans’ package(*23*) in R).

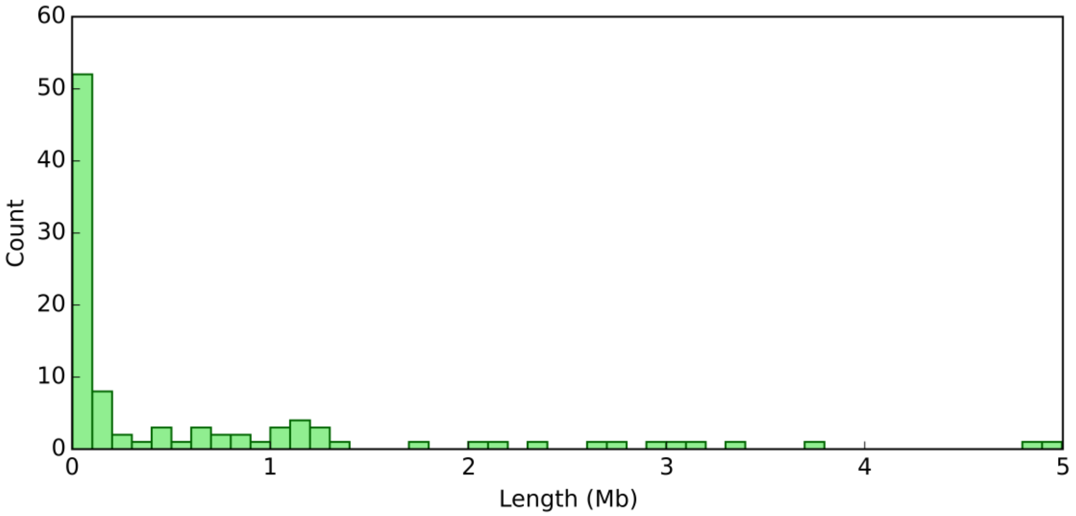

Fig. S2 Size distribution of the reconstructed haplotype blocks (selected alleles) in hot-evolved replicates. 50% of the haplotype blocks were smaller than 100Kb but ~25% were larger than 1Mb.

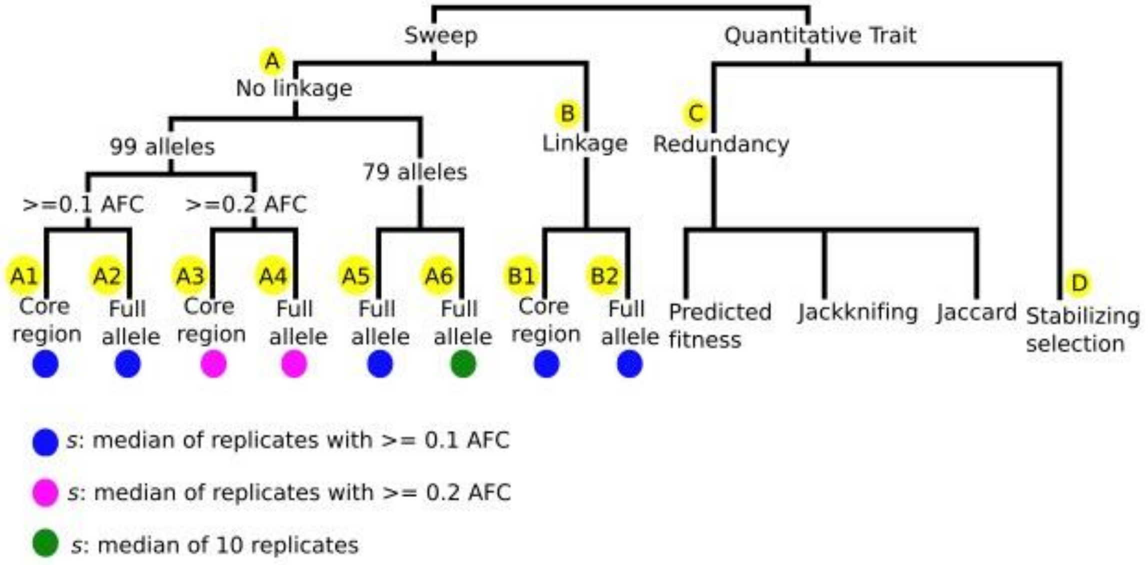

Fig. S3 Different simulation scenarios used to contrast selective sweep and quantitative trait models. We compare different adaptive sweep and quantitative trait scenarios to the empirical data: selective sweep models of independent (A) and linked alleles (B) were studied as well as different aspects of a quantitative trait model: genetic redundancy (C) and simulations of allele frequency changes assuming a quantitative trait with stabilizing selection (D). Sweep models (A and B) were performed for 99 (A1-A4, B1, B2) and 79 alleles (increasing in more than four replicates, A5, A6). The selection coefficient (*s*) was estimated using the median frequency trajectories of selected alleles in replicates with ≥ 0.1 (blue circles) and ≥ 0.2 (pink circles) frequency increase. *s* was also estimated using the median frequency change in all 10 replicates without conditioning on a frequency increase of at least 0.1 (green circle). *s* and starting frequency of the selected alleles were estimated using either the ‘full alleles’ or only the ‘core region’ of selected alleles (see ‘Different approaches to estimate allele frequencies for inference of selection coefficient’ for definition of core region). The details of the redundancy model are explained in SI Methods: ‘C. Genetic redundancy model’. Simulations of a quantitative trait with stabilizing selection were performed with 99 loci using starting frequency of selected alleles (‘full allele’) and equal effect sizes of all alleles.

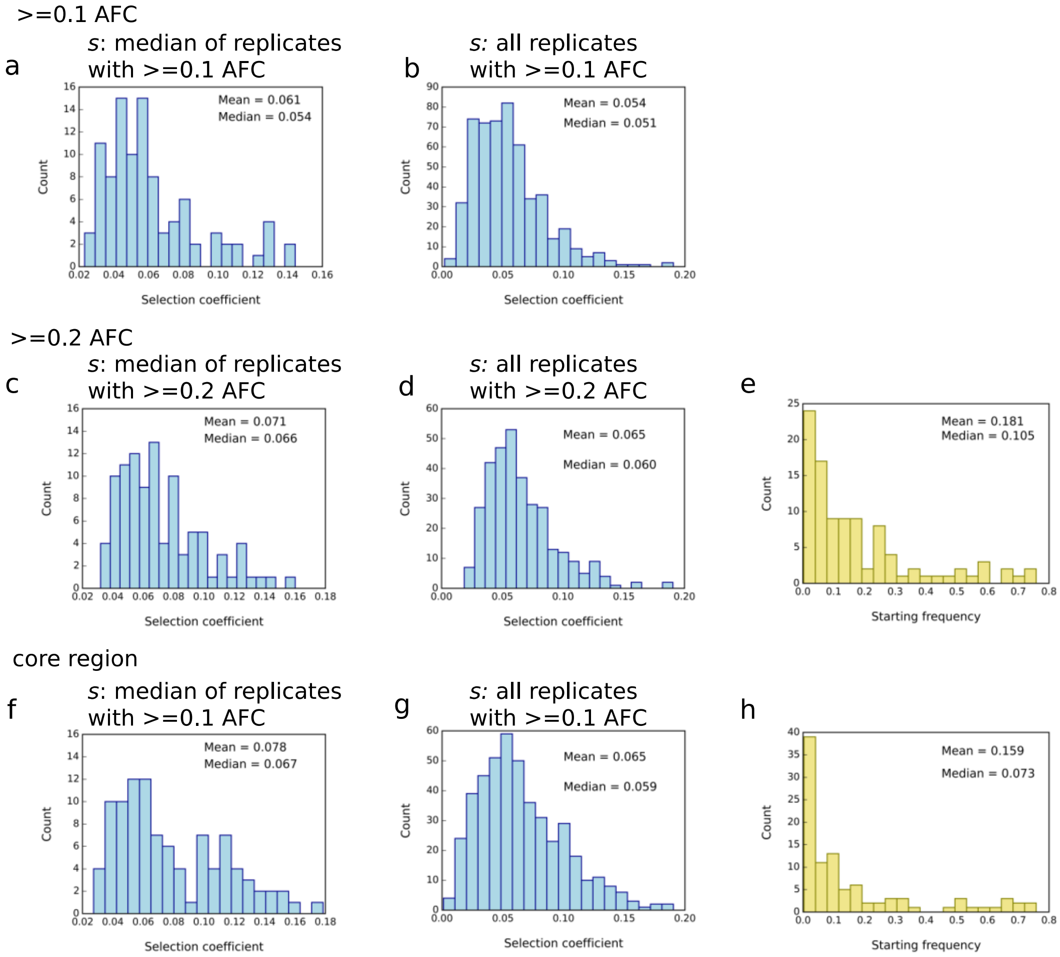

Fig. S4 Selection coefficients (*s*) of selected alleles using different approaches to estimate the frequency of a given selected allele. The median frequency of each allele (the median frequency of all SNPs characteristic to a haplotype block/allele) was computed and the frequency trajectory of replicates with ≥0.1 and ≥0.2 frequency increase until generation 60 were used for *s* estimation (0.1 AFC panel: A-B, 0.2 AFC panel: C-E). The core region panel (F-H) shows *s* for the region with the highest estimated *s* in each allele (see 'Different approaches to estimate allele frequencies for inference of selection coefficient' for definition of core region). *s* was estimated separately for each replicate with ≥0.1/ ≥0.2 AFC and the median *s* across the replicates is reported in A,C,F) whereas in B,D,G) the calculated *s* for all the replicates with ≥0.1/ ≥0.2 AFC is reported. The starting frequency of alleles with ≥0.1 and ≥0.2 frequency increase (E) and the core regions of selected alleles (H) is shown. The estimated *s* using all approaches agree (similar mean and median), but frequency trajectories of replicates with ≥0.1 AFC resulted in the most conservative *s* estimation.

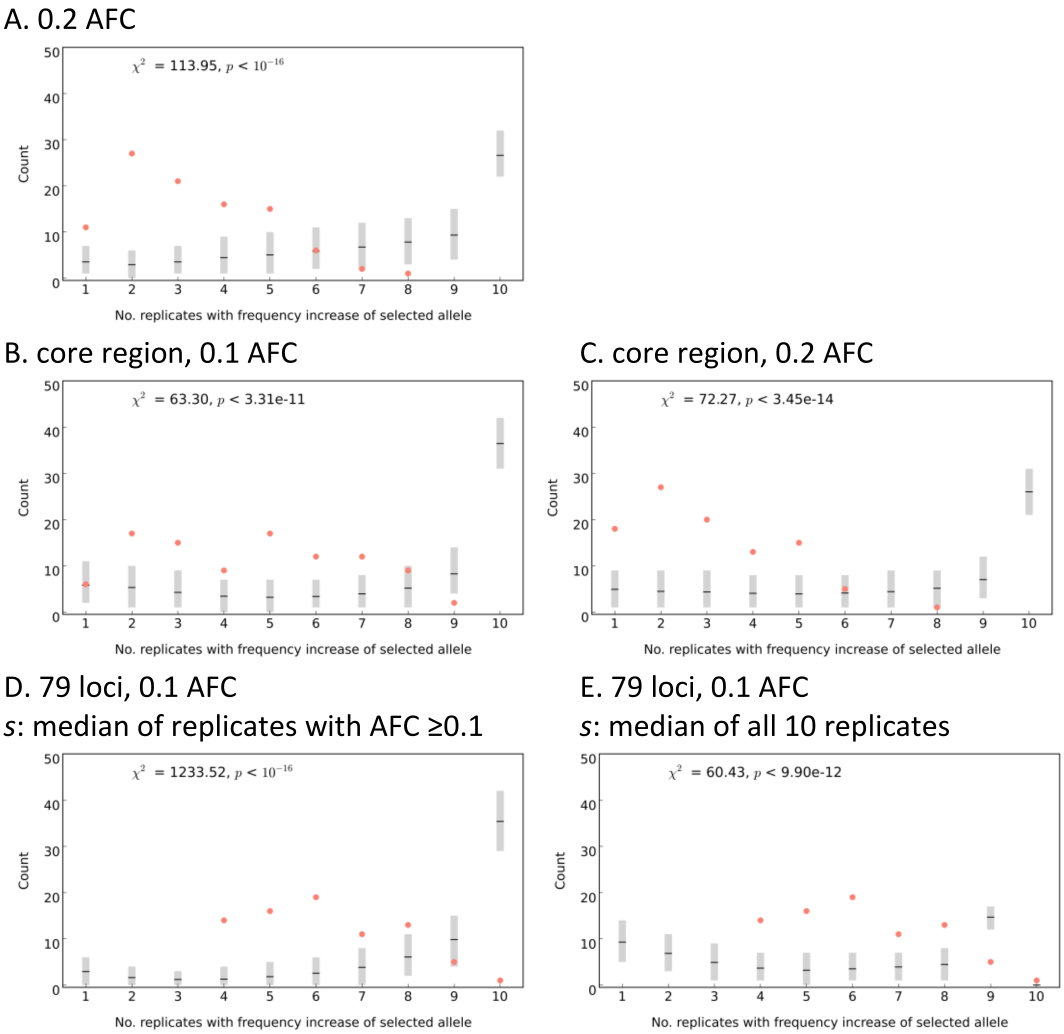

Fig. S5 Simulations to test for a sweep model without linkage and a constant selection coefficient across replicates. The frequency distribution of replicates in which selected alleles (99 alleles except 79 alleles in D and E) increase in frequency (B, D, E; ≥0.1, A, C: ≥0.2) until generation 60. Observed data are indicated by salmon dots. The expected distribution assuming constant *s* across replicates and time and independence of alleles was obtained by computer simulations (see 'Simulation to test for a sweep model without linkage and a constant *s* across replicates and time') and is indicated in gray (mean in black). A) simulations performed with *s* estimated from frequency trajectories of replicates with AFC ≥0.2 (Fig. S4C), B,C) simulations performed with *s* estimated for the core region of each selected allele using frequency trajectories of replicates with AFC ≥0.1 (B, Fig. S4F) and ≥0.2 (C), D,E) simulations performed using only alleles that increased in frequency (≥0.1) in ≥4 replicates. Note that alleles identified in only 1-3 replicates had high *s* and low starting frequency and were therefore excluded from these simulations. D) *s* is estimated by median *s* among replicates with AFC ≥0.1. E) simulations are based on *s* estimated from the median across all replicates (even if alleles have < 0.1 AFC). Starting frequencies of simulated alleles match the empirical data (Fig. S4E, H). All simulations assume free recombination and no linkage among loci.

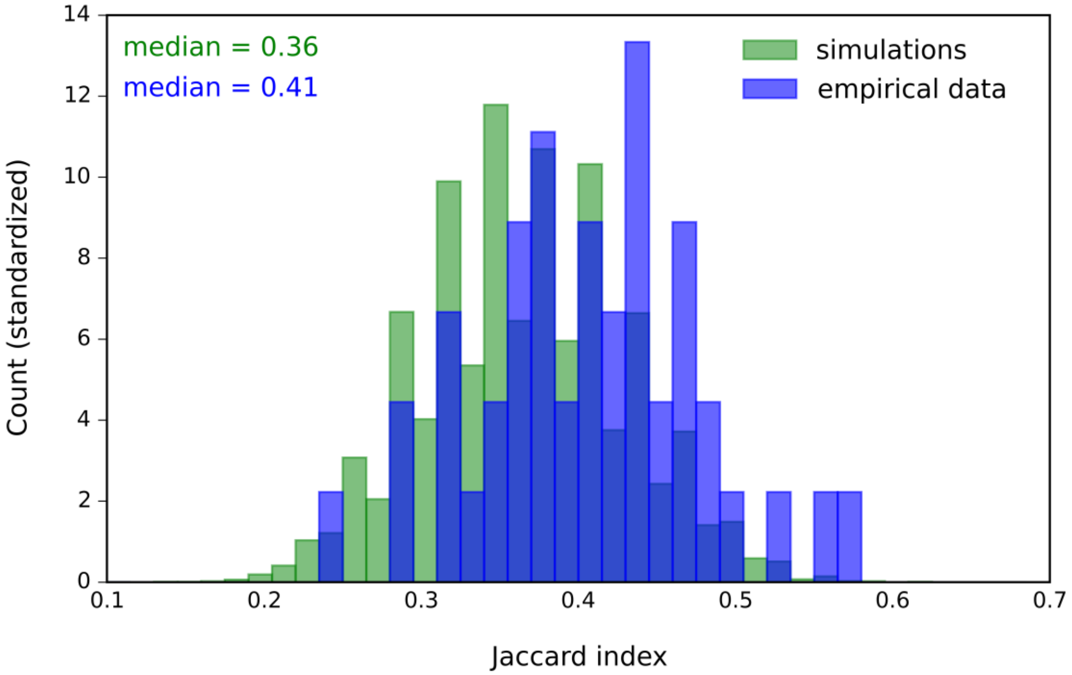

Fig. S6 Hot-evolved replicates are significantly more similar than expected by chance. Distribution of pairwise Jaccard indices among hot-evolved replicates (blue) and randomly combined alleles (green, redundancy model Fig. 4B). The counts in each data set have been standardized to sum up to 100. The median Jaccard index among hot-evolved replicates (0.41) is significantly (*p*<0.001) higher than the Jaccard indices of randomly combined alleles (90 percentile = 0.37).

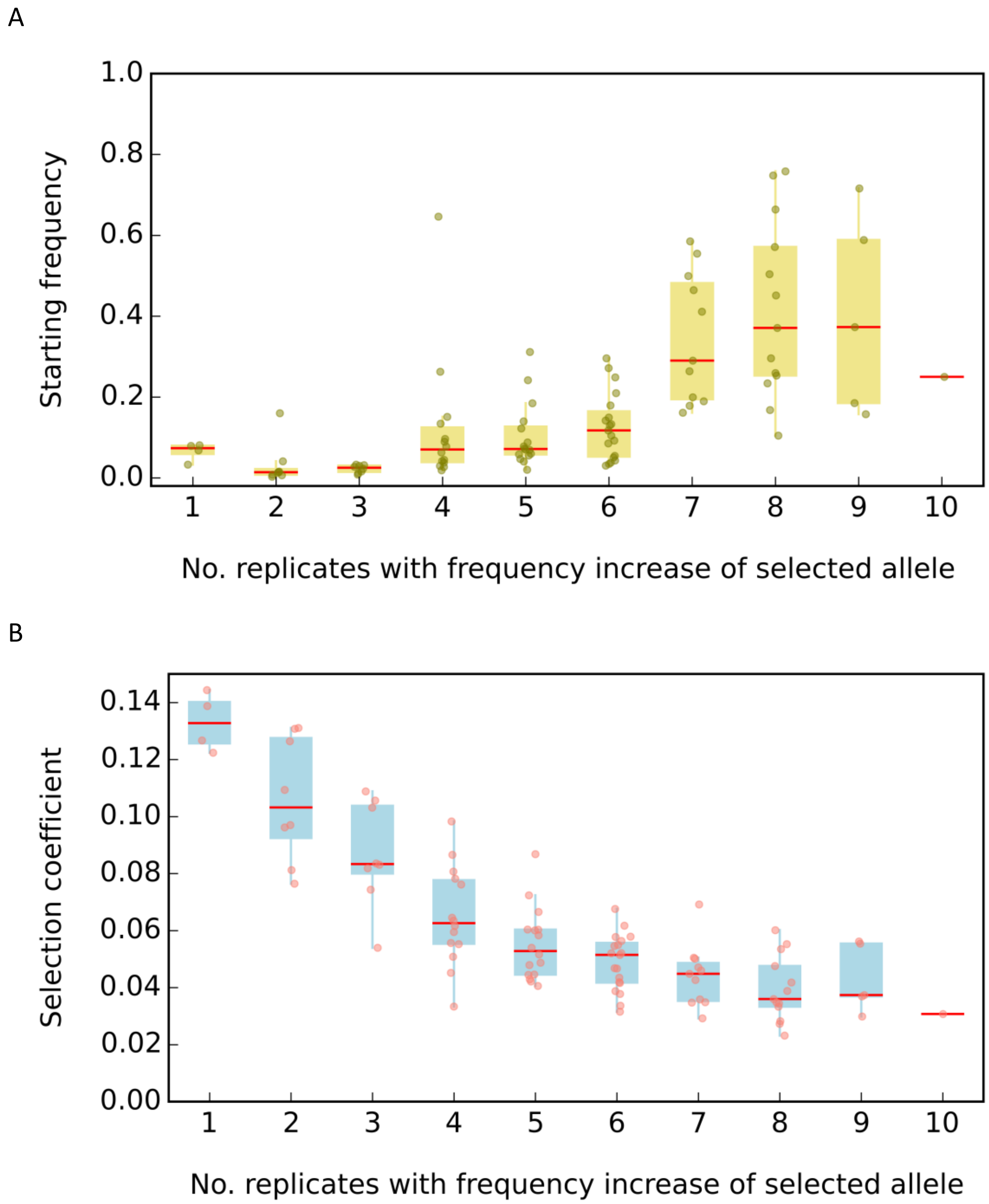

Fig. S7 Characteristics of selected alleles. A) starting frequency and B) selection coefficient (s) of the selected alleles classified by the number of replicates in which a given selected allele has ≥0.1 frequency increase at generation 60. The selected alleles which increased in frequency (≥0.1) in only one to three replicates have low starting frequencies and the highest estimated s. Boxplots show the 1st and 3rd quartile of the distribution, and horizontal bars in each box shows the median in each category. The data of individual selected alleles are shown as scattered dots in each boxplot.

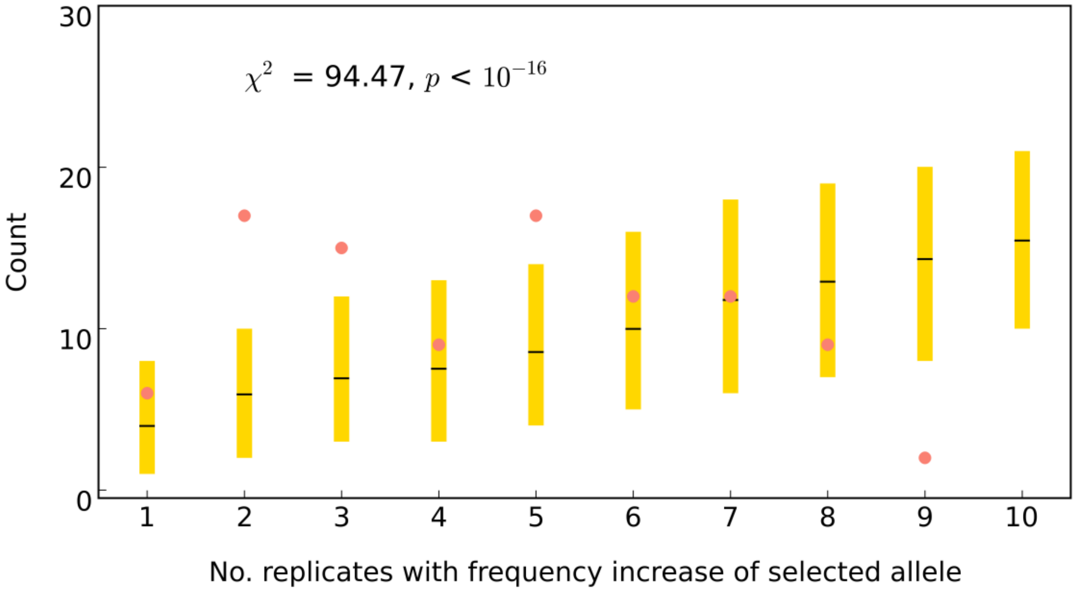

Fig. S8 Testing a sweep model with linkage and a constant selection coefficient across replicates. The frequency distribution of replicates in which selected alleles increase ≥0.1 in frequency at generation 60. Observed data are indicated by salmon dots. The expected distribution based on constant *s* across replicates and time was obtained by computer simulations which account for linkage (Materials and Methods) and is indicated in golden bars (mean in black). Simulations were performed using the starting frequency (Fig. S4H) and the estimated *s* (Fig. S4F) of the core region of selected alleles.

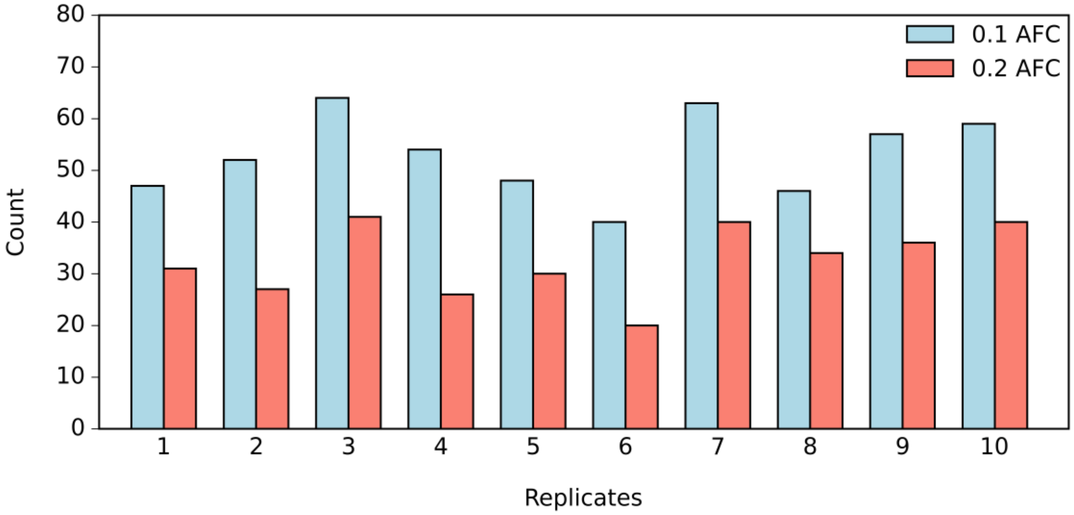

Fig. S9 Frequency of the selected alleles in hot-evolved replicates. The number of selected alleles with at least 0.1 allele frequency change (0.1 AFC) at generation 60 varies among replicates (average 53 loci). The difference among replicates still persists when selected alleles with at least 0.2 allele frequency change (0.2 AFC) at generation 60 (average 32.5) are considered.

SI Tables

Table S1 Characteristics of the reconstructed haplotype blocks.

# This table is provided as an additional Excel file

Table S2 Enrichment of gene functions in selected alleles.

# This table is provided as an additional Excel file

Table S3 Enrichment of KEGG pathways in selected alleles.

# This table is provided as an additional Excel file

Table S4 Absolute frequency difference of the identified P-elements in selected alleles.

Chromosome	Number	Delta
2	1	0.1778
2	8	0.325
2	17	0.1854
2	19	0.1338
2	21	0.4038
3	3	0.1259
3	16	0.205
3	29	0.3062
3	30	0.465
3	41	0.4258

Chromosome: left and right arms of chromosomes are concatenated as some haplotype blocks span the centromere. Number: an arbitrary number given to the haplotype block of each chromosome (similar to the numbers in Table S1). Delta: the absolute frequency difference between the frequency of the allele in any of the two replicates with highest frequency at generation 60 and the frequency of the P-element in those replicates.

Table S5 Estimated *N_e* in hot-evolved replicates for autosomes and the X chromosome.

Replicate	Autosomes	X chromosome
1	381	217
2	335	280
3	247	269
4	255	240
5	246	208
6	307	274
7	297	270
8	244	241
9	310	358
10	287	268
Mean	**291**	**262**

Table S6 The coverage of SNPs for all time points and replicates.

# This table is provided as an additional Excel file