Low concordance of short-term and long-term selection responses in experimental *Drosophila* populations

Anna Maria Langmüller[1,2] and Christian Schlötterer[2,*]

[1]Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, 1210 Wien, Austria

[2]Vienna Graduate School of Population Genetics, Vienna, Austria

[*]corresponding author: christian.schloetterer@vetmeduni.ac.at

February 6, 2020

Abstract

Experimental evolution is becoming a popular approach to study the genomic selection response of evolving populations. Computer simulation studies suggested that the accuracy of the signature increases with the duration of the experiment. Since some assumptions of the computer simulations may be violated, it is important to scrutinize the influence of the experimental duration with real data. Here, we use a highly replicated Evolve and Resequence study in *Drosophila simulans* to compare the selection targets inferred at different time points. At each time point, approximately the same number of SNPs deviated from neutral expectations, but only 10 % of the selected haplotype blocks identified from the full data set could be detected in the first 20 generations. Those haplotype blocks that emerged already after 20 generations differ from the others by being strongly selected at the beginning of the experiment and display a more parallel selection response. Consistent with previous computer simulations, our results demonstrate that only Evolve and Resequence experiments with a sufficient number of generations can characterize complex adaptive architectures.

KEYWORDS: experimental evolution, Evolve & Resequence, *Drosophila simulans*, early adaptation, replicated time series data, concordance of adaptation

Introduction

Deciphering the adaptive architecture has been of long-standing interest in evolutionary biology. In contrast to natural populations, experimental evolution (EE) provides the possibility to replicate experiments under controlled, identical conditions and to study how evolution shapes populations in real time (Kawecki *et al.* (2012)). The combination of EE with next-generation sequencing - Evolve and Resequence (E&R) (Turner *et al.* (2011); Long *et al.*

(2015); Schlötterer *et al.* (2015)) - has become a popular approach to study the genomic response to selection and to identify adaptive loci. E&R has been applied to various selection regimes, such as virus infection (Martins *et al.* (2014)), host-pathogen co-adaptation (Papkou *et al.* (2019)), thermal adaptation (Orozco-Terwengel *et al.* (2012); Barghi *et al.* (2019)), or body weight (Johansson *et al.* (2010)). A wide range of experimental designs have been used, which vary in census population size, replication level, history of the ancestral populations, selection regime, and number of generations (Garland and Rose (2009); Turner *et al.* (2011); Kawecki *et al.* (2012); Lang *et al.* (2013); Burke *et al.* (2014); Huang *et al.* (2014); Hardy *et al.* (2018); Castro *et al.* (2019); Michalak *et al.* (2019); Seabra *et al.* (2019)). The duration of E&R studies ranged from less than 20 generations (Turner and Miller (2012); Kelly and Hughes (2018); Rêgo *et al.* (2019)), to a few dozen (Johansson *et al.* (2010); Orozco-Terwengel *et al.* (2012)) and even hundreds of generations (Burke *et al.* (2010)). Computer simulations showed that the number of generations has a strong influence on the power of E&R studies in sexually reproducing organisms, and increasing the number of generations typically increased it (Baldwin-Brown *et al.* (2014); Kofler and Schlötterer (2014); Vlachos and Kofler (2019)). A larger number of generations increases the opportunity for more pronounced allele frequency changes and more recombination events, which results in a more refined mapping resolution (Baldwin-Brown *et al.* (2014); Kofler and Schlötterer (2014); Kessner and Novembre (2015)). Since simulations make simplifying assumptions, it is important to scrutinize their conclusions with empirical data. Until recently no suitable data-sets of obligate outcrossing populations were available, which included multiple time points, were replicated and had starting allele frequencies matching natural populations. We use an E&R experiment (Barghi *et al.* (2019)), which reports allele frequency changes in 10 replicates over 60 generations in 10 generation intervals, to investigate the impact of the experimental duration on the observed genomic response. The time resolved genome-wide polymorphism data of this experiment allow to contrast putative selection targets which are inferred at different time

points. We show that only a subset of the selection targets are detected at earlier generations, which are not representative of the underlying adaptive architecture.

Material & Methods

Experimental *Drosophila simulans* populations

A detailed description of the *Drosophila simulans* E&R experiment can be found in Barghi *et al.* (2019) and Hsu *et al.* (2019). Pooled individuals (Schlötterer *et al.* (2014)) from the evolving populations were sequenced every 10th generation starting with the founder population (generation 0) until generation 60. This E&R experiment started from 202 isofemale lines, which were collected in Tallahassee, Florida. 10 replicate populations evolved in the laboratory at a "cycling hot" temperature regime (12 hours light and 28 °C, 12 hours dark and 18 °C). The census size of the replicates was 1 000 individuals with non-overlapping generations (Barghi *et al.* (2017, 2019); Hsu *et al.* (2019)).

Genomic analysis hierarchy

We investigated the genomic response of the experimental *Drosophila* populations on three different levels: candidate SNPs, candidate SNPs in a window with a fixed number of SNPs, and candidate SNPs shared with reconstructed selected haplotype blocks. A detailed description for each level is given below, and the different hierarchies are depicted in Figure 1. Reasoning that the most reliable signal is detected at the most advanced generation (60), we performed the same analysis at earlier time points and determined to what extent the same selection targets were identified as in generation 60.

Identification of candidate SNPs

Barghi *et al.* (2019) applied various filtering steps to obtain a robust SNP set from the ancestral population. In short, SNPs were called applying the following criteria: base quality of 40 in at least one replicate, a coverage between the 2nd and 98th percentile, and the minor allele is supported by at least ten reads. Repeats, transposable elements, SNPs specific to Y-translocated genes (Tobler *et al.* (2017)), and 5-bp regions around indels were excluded from the analysis to increase the robustness of the SNP set (for further details, see Barghi *et al.* (2019)), resulting in 5,096,200 SNPs on chromosomes X, 2, 3, and 4.

To study the selection response at different time points, we identified "candidate SNPs" based on the frequency difference between the ancestral and evolved populations for each of the six time points. Following Barghi *et al.* (2019), replicates were tested separately (Fisher's exact test) and jointly (Cochran-Mantel-Haenszel test, CMH) to identify SNPs with pronounced allele frequency change using `PoPoolation2` (Kofler *et al.* (2011)). We chose the CMH test statistic out of a range of approaches that allow to identify selected SNPs. It has been shown recently that the CMH test – although not taking intermediate time points into account – consistently outperforms other methods regardless of the investigated selection scenario (Vlachos *et al.* (2019)). Minimum and maximum coverage restrictions were not imposed because outlier SNPs with extreme coverage had already been removed in the original study. Neither the CMH test nor the Fisher's exact test account for allele frequency change due to genetic drift. To detect SNPs that show more allele frequency change than expected under drift, we performed neutral simulations with `Nest` (Jónás *et al.* (2016)) using estimates of the effective population size (N_e) between generation 0 and the focal time point (Table S1–S3). The neutral simulations further used the empirical starting allele frequencies and sequencing coverages. For the CMH test, N_e estimates were averaged across replicates for autosomes and the X chromosome separately. For Fisher's exact test, we used replicate-specific N_e estimates of the autosomes. Based on these neutral simulations we determined candidate SNPs

with a false discovery rate smaller than 5 % (Barghi *et al.* (2019)).

We identified 56,166 candidate SNPs in generation 60, compared to 55,199 in Barghi *et al.* (2019). This small discrepancy can be explained by stochastic differences arising from the neutral simulations used to determine the significance threshold. We excluded six out of 99 haplotype blocks from Barghi *et al.* (2019) with less than 90% of the previously reported candidate SNPs.

Identification of candidate windows

The number of candidate SNPs is inflated as a result of linkage disequilibrium in the experimental populations (Nuzhdin and Turner (2013); Tobler *et al.* (2014)). To account for non−independence of candidate SNPs, we used a window-based approach. We split the main chromosomes (X, 2 and 3) into non-overlapping windows of 5 000 SNPs that are segregating in all generations and replicates. We chose SNPs instead of base pairs as window size measure to allow for variation in SNP density along the genome. To determine if a given window contains more candidate SNPs than expected, we sampled the same number of random SNPs as candidate SNPs in this window (1 000 iterations) onto the whole genome. "Candidate windows" contained at least as many candidate SNPs as the 99th percentile of randomly sampled SNPs. Applying the procedure independently to candidate SNPs from all time points provides time point specific candidate windows (Figure S1). Varying the window size of 5 000 segregating SNPs or allowing SNPs to fix resulted in qualitatively similar results (Table S4 , S5). We evaluated the similarity of two time points with the Jaccard index - a dimensionless parameter ranging from 0 (no overlap) to 1 (sets are identical) - for both candidate SNPs, and candidate windows.

The number of candidate SNPs in a window is a summary statistic which ignores the significance of the candidate SNPs. If a signal is robust between two time points, we expect the same p-value-based ranking of candidate SNPs assuming homogeneous coverage for all sites

and time points. Although coverage heterogeneity among sites will change the confidence in the estimates of the allele frequency change, the average dynamics of relative significance allow us to determine whether the robustness of a putative selection response increases with time. Thus, we also evaluated whether candidate SNPs in a given window had a similar significance rank. For each candidate window we created a ROC-like curve (similar to Jakšić and Schlötterer (2016)) by ranking the candidate SNPs by their p-values - the most significant SNP was assigned rank 1 - and calculating the overlap in top-ranked SNPs between two time points.

HAplotype block Discovery Rate (HADR)

Barghi *et al.* (2019) clustered candidate SNPs from generation 60 into selected haplotype blocks based on similar allele frequency trajectories over time and replicates (Franssen *et al.* (2016)). The reconstructed haplotype blocks were further validated with experimentally phased haplotypes from ancestral and evolved populations (Barghi *et al.* (2019)) and 96 % of the reconstructed haplotype blocks could be confirmed by the experimentally derived haplotypes. This suggests that reconstructed haplotype blocks provide a robust set of linked candidate SNPs.

Taking advantage of this additional confirmation of the candidate SNPs in a selected haplotype block, we developed a third measure of similarity between time points. We determined the fraction of candidate SNPs comprising a haplotype block that were also discovered at a given time point (haplotype block discovery rate, HADR) using the `poolSeq` R-package (Taus *et al.* (2017)). We note that inference of selected haplotype blocks at each generation does not provide a good alternative to HADR, as the ability to cluster SNPs into haplotype blocks is dependent on the number of time points (Franssen *et al.* (2016)), resulting in less power at early time points compared to later ones.

Early Detected HAplotype blocks (EDHAs)

We applied hierarchical clustering (Pollard and Laan (2005)), PCA and kmeans (Hartigan and Wong (1979)) to group haplotype blocks based on their HADR patterns. The hyper-parameter k, which determines the number of clusters, was set to 5 based on the gap statistic (Tibshirani *et al.* (2001))and k-means clustering resulted into a group of 10 haplotype blocks with elevated HADR in generation 20 (Figure S2). This group of 10 haplotype blocks can also be separated from other haplotype blocks by the first principal component of a PCA applied to HADR from generation 10 to generation 50 (Figure S3). We refer to the haplotype blocks in this cluster as early detected haplotype blocks (EDHAs). We evaluated whether EDHAs have distinct characteristics compared to all other haplotype blocks using the following features: haplotype block length, median starting allele frequency, average recombination rate (Howie *et al.* (2019)), selection coefficient (s, estimated with `poolSeq` (Taus *et al.* (2017))) in generation 20, s in generation 60, selection coefficient ratio $r_s = \frac{s_{20}}{s_{60}}$, and number of rising replicates in generation 20. Following Barghi *et al.* (2019) , we classified a haplotype block as replicate specific, if the allele frequency of candidate SNPs from a haplotype block increases on average by at least 10%. To test the robustness of the replicate specific rising behaviour, we also used 5, 15, and 20% thresholds. Selection coefficients were averaged over replicates that passed the allele frequency change threshold.

Simulation of an E&R experiment

Despite the selected haplotype blocks were inferred with high confidence by Barghi *et al.* (2019), we were interested to confirm our conclusions with a simulated data set for which the selection targets are known. Because the distinction between true causative SNPs and linked polymorphisms is challenging (Nuzhdin and Turner (2013); Tobler *et al.* (2014)), we obtained a simulated data set that models linkage and resembles the Barghi *et al.* (2019) study using

MimicrEE2 (v. 206, Vlachos and Kofler (2018)). Our ancestral population was built from 189 ancestral haplotypes published in Barghi *et al.* (2019) and Howie *et al.* (2019). We simulated a selective sweep scenario with 99 independent selection targets, where each of them is located in the inferred selected haplotype block, has the same starting frequency, and selection coefficient (both estimated as the median from all selected SNPs in the block) (Barghi *et al.* (2019)). Since, we were primarily interested in the early generations, we reasoned that selective sweeps approximate the selection trajectories for a polygenic trait with stabilizing selection rather well (Chevin and Hospital (2008); Hayward and Sella (2019)). Using the *Drosophila simulans* recombination rate map (Howie *et al.* (2019)), we simulated 10 outcrossing populations with constant census size adapting for 60 non-overlapping generations. We sampled the read depth for each site from a Poisson distribution with λ equal to the average coverage for each population and time point as reported in Barghi *et al.* (2019). For each site, we applied binomial sampling with the determined coverage and true allele frequency to mimic the sampling of reads out of a DNA pool (Jónás *et al.* (2016)). The identification of candidate SNPs and windows of the simulation followed the protocol of the empirical data (as described above). In contrast to the empirical data, the causative SNPs are known. Thus, instead of relying on the dynamics of selected haplotype blocks (see HADR), we investigated the number of selection targets that can be found amongst statistical outliers across generations. Early detected targets (see EDHAs) were defined as targets that are consistently detected as outlier from generation 20 onwards.

Results & Discussion

Subsequent time points are more similar for advanced generations

We studied the similarity of selection signatures for different time points using 10 replicates of a *D. simulans* population, which evolved for 60 generations to a novel hot environment

(Barghi *et al.* (2019)). With Pool-Seq (Schlötterer *et al.* (2014)) data from every 10th generation, we evaluated the selection signature on three different levels: candidate SNPs, candidate SNPs in a window spanning a fixed number of SNPs, and candidate SNPs shared with reconstructed selected haplotype blocks (Figure 1). The similarity of inference between two time points was determined by the Jaccard index (J), ranging from 0 (no overlap between two SNP sets) to 1 (sets are identical). We found that all candidate SNP sets are more similar than expected by chance. The Jaccard index ranged from 0.08 (generation 10 vs generation 60) to 0.40 (generation 50 vs generation 60), where subsequent time points are more similar than those separated for more than 10 generations (e.g. J=0.15 (generation 10 vs generation 20); J=0.08 (generation 10 vs generation 60)). Furthermore, the similarity of candidate SNP sets from subsequent time points increases with time until it ultimately more than doubles for the last two generations (J=0.15 (generation 10 vs generation 20); J=0.34 (generation 50 vs generation 60), Figure 2A). The monotonic increase in similarity with time shows that more similar selection patterns are detected at later generations. While this suggests that outliers at generation 60 provide the most reliable selection signature, we were interested to confirm this with computer simulations where all selection targets and the type of selection are known. We simulated an E&R experiment that resembles the data from Barghi *et al.* (2019) and repeated our analysis on simulated, time resolved allele frequencies. Again, we observe that candidate SNP sets are more similar than expected by chance, the majority of subsequent time points are more similar than those separated for more than 10 generations, and the similarity of candidate SNP sets from subsequent time points increases with time (Figure S4).

Since the analysis of single SNPs suffers from considerable stochasticity, and neighboring SNPs are not independent (Tobler *et al.* (2014); Howie *et al.* (2019)), we repeated the analysis of different time points using non-overlapping windows of 5 000 SNPs. Reasoning that windows containing a target of selection will harbor multiple candidate SNPs, we defined

selected windows as those, which harbor more candidate SNPs than expected by chance. Consistent with higher stochasticity at the SNP level, a higher similarity was observed for candidate windows (from J=0.26 (generation 10 vs generation 60) to J=0.62 (generation 40 vs generation 50)). Again, adjacent time points have a higher Jaccard index than time points farther apart (J=0.26 (generation 10 vs generation 60); J=0.39 (generation 10 vs generation 20)). The similarity of subsequent time points also increases with the duration of the experiment (J=0.39 (generation 10 vs generation 20) ; J=0.59 (generation 50 vs generation 60), Figure 2A). In contrast to the SNP level, the set of selected windows after 10 generations is only significantly similar to generation 20, but not to any other generation. Thus, the pattern of less reliable selection targets in the early generations is confirmed at the window level, albeit with different significance levels. The same pattern was noticed for the simulated E&R data (Figure S4).

For an alternative measure of similarity, we used the ranking of candidate SNPs in a specific window based on their p-values and compared it between different time points. If a signal is robust between two time points, we expect the same SNP ranking in a selected window. Consistent with the other tests, we found that the congruence in candidate SNP ranking increases with time in both empirical and simulated data (Figure 3, S5). To rule out that rare SNPs are responsible for the dissimilarity between early and late time points, we calculated similarity measures based on SNPs that are segregating at all generations and time points. Nevertheless, including SNPs which were lost in at least one replicate during the experiment did not result in a pronounced decrease in similarity (Figure S6).

The analysis of selected haplotype blocks provides another possibility to control for non-independence of single candidate SNPs. We calculated the haplotype block discovery rate (HADR) - the fraction of candidate SNPs in a haplotype block that are rediscovered at a given time point. Similar to the other analyses, we observe higher similarity between later time points (Figure 2B), with a pronounced increase of median HADR between generation 30 (<

25 %) and 40 (> 50 %). In the simulated data, the number of detected targets monotonically increased until generation 50 (Figure S7).

Independent of the measure of similarity between time points, for both empirical and simulated data, the selection signatures of early time points are consistently less robust than those from later time points. This observation highlights that a more reliable identification of selection targets requires longer experiments with additional generations. With the limitation that for empirical data the true targets of selection are not known, in the Barghi *et al.* (2019) data a particularly striking observation is that a similar number of candidate SNPs was detected at each time point despite the number of "true" targets increased with the duration of the experiment (Table S6). While this suggests that earlier time points harbor more false positives, it may be possible that some targets were only selected during the first generations. It is important to note that the sweep simulations did not match this result, rather the number of false positives increased with the duration of the experiment (Table S7). We attribute this discrepancy of the simulated and empirical data outcome to the selection regime employed, which affects mainly later generations. Sweep simulations result in haplotype blocks with many selection targets (Barghi and Schlötterer (2019)), which in turn influences the number of candidate SNPs. Simulations of polygenic adaptation require many unknown parameters to match the empirical data (Hayward and Sella (2019)), which limits the power of computer simulations to scrutinize this result further. Rather, experimental validation of selection targets in secondary E&R studies (Burny *et al.* (2019)) may be used to confirm selection signatures beyond statistical testing and could serve an important role to test early selection signatures that cannot be confirmed at later time points.

Only few selection targets are shared across all generations

More than 27 000 candidate SNPs can be identified at each time point (Table S6), but only a small (4.8 %) subset is consistently detected at every generation (Figure 4; including rare

SNPs see Figure S8). Apart from highlighting the more robust selection signatures with an increasing number of generations, this analysis raises an important concern about the usefulness of meta-analyses on the SNP level. With less than 5% of the SNPs being shared in the same selection experiment, it will be extremely difficult to compare studies that started from different founder populations and were selected for a different number of generations.

We repeated the analysis for windows and determined the number of selected windows that are shared across all generations. With 18 out of 74 candidate windows in generation 60 (24.3 %, Figure 4) being detected at all generations, the window analysis shows more consistency across time points than a SNP-based analysis. This observation is independent of window size (Table S4, and S5) the inclusion of rare SNPs into the analysis (Figure S8). Because we observed a similar trend in our simulated experiment (Figure S9), we propose that meta-analyses of E&R data should be performed on the level of windows, or based on selected haplotype blocks to avoid false negatives due to the high stochasticity of SNP-based analyses.

Selection signatures detected early in the experiment are not representative of the underlying adaptive architecture

This study focused on the comparison of selection targets detected at early and late time points. Since analyses based on single SNPs are very stochastic, we investigated the fraction of candidate SNPs comprising a haplotype block that were also discovered at earlier time points (HADR). We detected 10 haplotype blocks with elevated HADR in generation 20 (early detected haplotype blocks, EDHAs, Figure S2, S3). We found that EDHAs do not differ in their starting allele frequency, haplotype block length, average recombination rate or absolute selection coefficients from other haplotype blocks (Figure S10). EDHAs are, however, more strongly selected at the beginning of the experiment, but are equally strongly selected as the remaining haplotype blocks at later generations (Figure 5A). Consistent with stronger

selection at earlier time points, the selection signature of EDHAs is significantly more parallel across replicates after 20 generations of adaptation in both empirical and simulated data. (Figure 5B, S11). We attribute this observation to a phenomenon similar to the "winners curse", that is that loci where stochastic effects increased the frequency in multiple replicates to enhance the contribution by selection are more likely to be detected.

All statistical tests, which are evaluating a parallel selection signature across replicates, are more likely to detect selection signatures shared across replicates, even with only moderate allele frequency changes. This could result in a biased picture of the underlying genetic architecture. The analysis of selection signatures in replicated experiments running for only a moderate number of generations is more likely to detect parallel than replicate specific selection signatures. This bias is not restricted to our study, but also an experimental study of *D. simulans* populations adapting 10 to 20 generations to a new temperature regime (Kelly and Hughes (2018)) found more parallel selection responses. We propose that additional analyses contrasting selection signatures of early and late time points are needed to confirm the enrichment of parallel selection signatures in short-term experiments.

ACKNOWLEDGMENTS

We thank the members of the Institut für Populationsgenetik for fruitful discussion and support. Special thanks to Neda Barghi, Sheng-Kai Hsu and Claire Burny for helpful comments on earlier versions of the manuscript. This work was supported by the Austrian Science Fund (FWF, grant number W1225) and an European Research Council (ERC) grant (ArchAdapt).

Bibliography

Baldwin-Brown, J. G., Long, A. D., and Thornton, K. R. 2014. The power to detect quantitative trait loci using resequenced, experimentally evolved populations of diploid, sexual organisms. *Molecular Biology and Evolution*, 31(4): 1040–1055.

Barghi, N. and Schlötterer, C. 2019. Distinct patterns of selective sweep and polygenic adaptation. *bioRxiv*.

Barghi, N., Tobler, R., Nolte, V., and Schlötterer, C. 2017. *Drosophila simulans*: A species with improved resolution in evolve and resequence studies. *G3: Genes, Genomes, Genetics*, 7(7): 2337–2343.

Barghi, N., Tobler, R., Nolte, V., Jakšić, A. M., Mallard, F., Otte, K. A., Dolezal, M., Taus, T., Kofler, R., and Schlötterer, C. 2019. Genetic redundancy fuels polygenic adaptation in *Drosophila*. *PLOS Biology*, 17(2): 1–31.

Burke, M. K., Dunham, J. P., Shahrestani, P., Thornton, K. R., Rose, M. R., and Long, A. D. 2010. Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature*, 467(7315): 587–590.

Burke, M. K., Liti, G., and Long, A. D. 2014. Standing genetic variation drives repeatable experimental evolution in outcrossing populations of *Saccharomyces cerevisiae*. *Molecular Biology and Evolution*, 31(12): 3228–3239.

Burny, C., Nolte, V., Nouhaud, P., Dolezal, M., and Schlötterer, C. 2019. Secondary evolve and re-sequencing: an experimental confirmation of putative selection targets without phenotyping. *bioRxiv*.

Castro, J. P., Yancoskie, M. N., Marchini, M., Belohlavy, S., Hiramatsu, L., Kučka, M., Beluch, W. H., Naumann, R., Skuplik, I., Cobb, J., Barton, N. H., Rolian, C., and Chan, Y. F. 2019. An integrative genomic analysis of the longshanks selection experiment for longer limbs in mice. *eLife*, 8: e42014.

Chevin, L. M. and Hospital, F. 2008. Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genetics*, 180(3): 1645–1660.

Franssen, S. U., Schlötterer, C., and Barton, N. H. 2016. Reconstruction of Haplotype-Blocks Selected during Experimental Evolution. *Molecular Biology and Evolution*, 34(1): 174–184.

Garland, T. and Rose, M. 2009. *Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments*.

Hardy, C. M., Burke, M. K., Everett, L. J., Han, M. V., Lantz, K. M., and Gibbs, A. G. 2018. Genome-Wide Analysis of Starvation-Selected *Drosophila melanogaster*-A Genetic Model of Obesity. *Molecular Biology and Evolution*, 35(1): 50–65.

Hartigan, J. A. and Wong, M. A. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1): 100–108.

Hayward, L. K. and Sella, G. 2019. Polygenic adaptation after a sudden change in environment. *bioRxiv*.

Howie, J. M., Mazzucco, R., Taus, T., Nolte, V., and Schlötterer, C. 2019. DNA Motifs Are Not General Predictors of Recombination in Two *Drosophila* Sister Species. *Genome Biology and Evolution*, 11(4): 1345–1357.

Hsu, S.-K., Jaksic, A. M., Nolte, V., Barghi, N., Mallard, F., Otte, K. A., and Christian, S. 2019. A 24 h Age Difference Causes Twice as Much Gene Expression Divergence as 100 Generations of Adaptation to a Novel Environment. *Genes*, 10(2).

Huang, Y., Wright, S. I., and Agrawal, A. F. 2014. Genome-wide patterns of genetic variation within and among alternative selective regimes. *PLOS Genetics*, 10(8): 1–15.

Jakšić, A. M. and Schlötterer, C. 2016. The interplay of temperature and genotype on patterns of alternative splicing in *Drosophila melanogaster*. *Genetics*, 204(1): 315–325.

Johansson, A. M., Pettersson, M. E., Siegel, P. B., and Carlborg, Ö. 2010. Genome-wide effects of long-term divergent selection. *PLOS Genetics*, 6(11): 1–12.

Jónás, Á., Taus, T., Kosiol, C., Schlötterer, C., and Futschik, A. 2016. Estimating the effective population size from temporal allele frequency changes in experimental evolution. *Genetics*, 204(2): 723–735.

Kawecki, T. J., Lenski, R. E., Ebert, D., Hollis, B., Olivieri, I., and Whitlock, M. C. 2012. Experimental evolution. *Trends in Ecology & Evolution*, 27(10): 547 – 560.

Kelly, J. K. and Hughes, K. A. 2018. Pervasive linked selection and intermediate-frequency alleles are implicated in an evolve-and-resequencing experiment of *Drosophila simulans*. *Genetics*, 211(3): genetics.301824.2018.

Kessner, D. and Novembre, J. 2015. Power analysis of artificial selection experiments using efficient whole genome simulation of quantitative traits. *Genetics*, 199(4): 991–1005.

Kofler, R. and Schlötterer, C. 2014. A guide for the design of evolve and resequencing studies. *Molecular Biology and Evolution*, 31(2): 474–483.

Kofler, R., Pandey, R. V., and Schlötterer, C. 2011. PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*.

Lang, G. I., Rice, D. P., Hickman, M. J., Sodergren, E., Weinstock, G. M., Botstein, D., and Desai, M. M. 2013. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, 500(7464): 571–574.

Long, A., Liti, G., Luptak, A., and Tenaillon, O. 2015. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nature Reviews Genetics*, 16(10): 567–582.

Martins, N. E., Faria, V. G., Nolte, V., Schlötterer, C., Teixeira, L., Sucena, É., and Magalhães, S. 2014. Host adaptation to viruses relies on few genes with different cross-resistance properties. *Proceedings of the National Academy of Sciences*, 111(16): 5938–5943.

Michalak, P., Kang, L., Schou, M. F., Garner, H. R., and Loeschcke, V. 2019. Genomic signatures of experimental adaptive radiation in *Drosophila*. *Molecular Ecology*, 28(3): 600–614.

Nuzhdin, S. V. and Turner, T. L. 2013. Promises and limitations of hitchhiking mapping. *Current Opinion in Genetics Development*, 23(6): 694 – 699. Genetics of system biology.

Orozco-Terwengel, P., Kapun, M., Nolte, V., Kofler, R., Flatt, T., and Schlötterer, C. 2012. Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Molecular Ecology*, 21(20): 4931–4941.

Papkou, A., Guzella, T., Yang, W., Koepper, S., Pees, B., Schalkowski, R., Barg, M.-C., Rosenstiel, P. C., Teotónio, H., and Schulenburg, H. 2019. The genomic basis of Red Queen dynamics during rapid reciprocal host–pathogen coevolution. *Proceedings of the National Academy of Sciences*, 116(3): 923–928.

Pollard, K. S. and Laan, M. J. V. D. 2005. Analysis of Genomic Data with Applications in R. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 167.

Rêgo, A., Messina, F. J., and Gompert, Z. 2019. Dynamics of genomic change during evolutionary rescue in the seed beetle *Callosobruchus maculatus*. *Molecular Ecology*, 28(9): 2136–2154.

Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. 2014. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, 15(11): 749–763.

Schlötterer, C., Kofler, R., Versace, E., Tobler, R., and Franssen, S. U. 2015. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. *Heredity*, 114(5): 431–440.

Seabra, S. G., Fragata, I., Antunes, M. A., Faria, G. S., Santos, M. A., Sousa, V. C., Simões, P., and Matos, M. 2019. Different Genomic Changes Underlie Adaptive Evolution in Populations of Contrasting History. *Molecular Biology and Evolution*, 36(6): 1358–1358.

Taus, T., Futschik, A., and Schlötterer, C. 2017. Quantifying Selection with Pool-Seq Time Series Data. *Molecular Biology and Evolution*, 34(11): 3023–3034.

Tibshirani, R., Walther, G., and Hastie, T. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B*, 63: 411–423.

Tobler, R., Franssen, S. U., Nolte, V., and Schlötterer, C. 2014. Patterns of Linkage Disequilibrium and Long Range Hitchhiking in Evolving Experimental *Drosophila melanogaster* Populations. *Molecular Biology and Evolution*, 32(2): 495–509.

Tobler, R., Nolte, V., and Schlötterer, C. 2017. High rate of translocation-based gene birth on the *Drosophila* Y chromosome. *Proceedings of the National Academy of Sciences*, 114(44): 11721–11726.

Turner, T. L. and Miller, P. M. 2012. Investigating natural variation in *Drosophila* courtship song by the evolve and resequence approach. *Genetics*, 191(2): 633–642.

Turner, T. L., Stewart, A. D., Fields, A. T., Rice, W. R., and Tarone, A. M. 2011. Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. *PLOS Genetics*, 7(3): 1–10.

Vlachos, C. and Kofler, R. 2018. Mimicree2: Genome-wide forward simulations of evolve and resequencing studies. *PLOS Computational Biology*, 14(8): 1–10.

Vlachos, C. and Kofler, R. 2019. Optimizing the power to identify the genetic basis of complex traits with Evolve and Resequence studies. *Molecular Biology and Evolution*.

Vlachos, C., Burny, C., Pelizzola, M., Borges, R., Futschik, A., Kofler, R., and Schlötterer, C. 2019. Benchmarking software tools for detecting and quantifying selection in evolve and resequencing studies. *Genome Biology*, 20(1): 169.

Data availability

Information about the data availability for the original report by Barghi *et al.* (2019) can be found under: https://doi.org/10.1371/journal.pbio.3000128

All scripts and simulations of this study will be available on GitHub upon publication.

Authors Contribution

CS designed the experiment. CS and AML designed the analysis. AML performed the bioinformatic analysis. CS and AML wrote the manuscript.

Tables and Figures

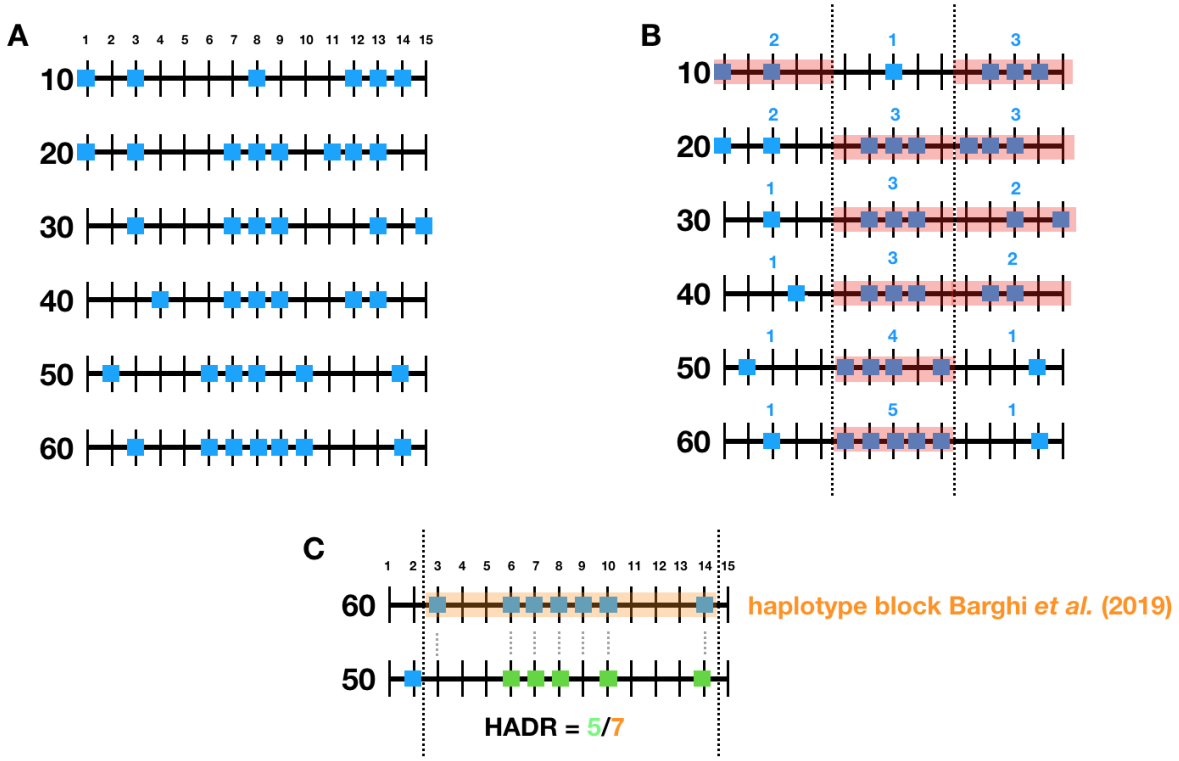

Figure 1: Genomic analysis hierachy: candidate SNPs (**A**), candidate SNPs in a window spanning a fixed number of SNPs (**B**), and candidate SNPs shared with reconstructed haplotype blocks (**C**).
(**A**): Candidate SNPs (blue squares) were determined for each generation and were compared either pairwise (Figure 2), or across multiple time points (Figure 4).
(**B**): Window-based approach to detect regions enriched for candidate SNPs (colored in red). The same number of random SNPs (red squares) as candidate SNPs were sampled onto the genome repeatedly (n=1 000) taking starting allele frequencies into account. Windows enriched for candidate SNPs contain at least as many candidate SNPs as the 99th percentile of the randomly sampled SNPs. Sets of candidate windows were either compared pairwise (Figure 2), or across multiple time points (Figure 4).
(**C**): Haplotype block discovery rate (HADR) = the fraction of candidate SNPs (blue squares in orange region) that were also discovered at a given time point (green squares).

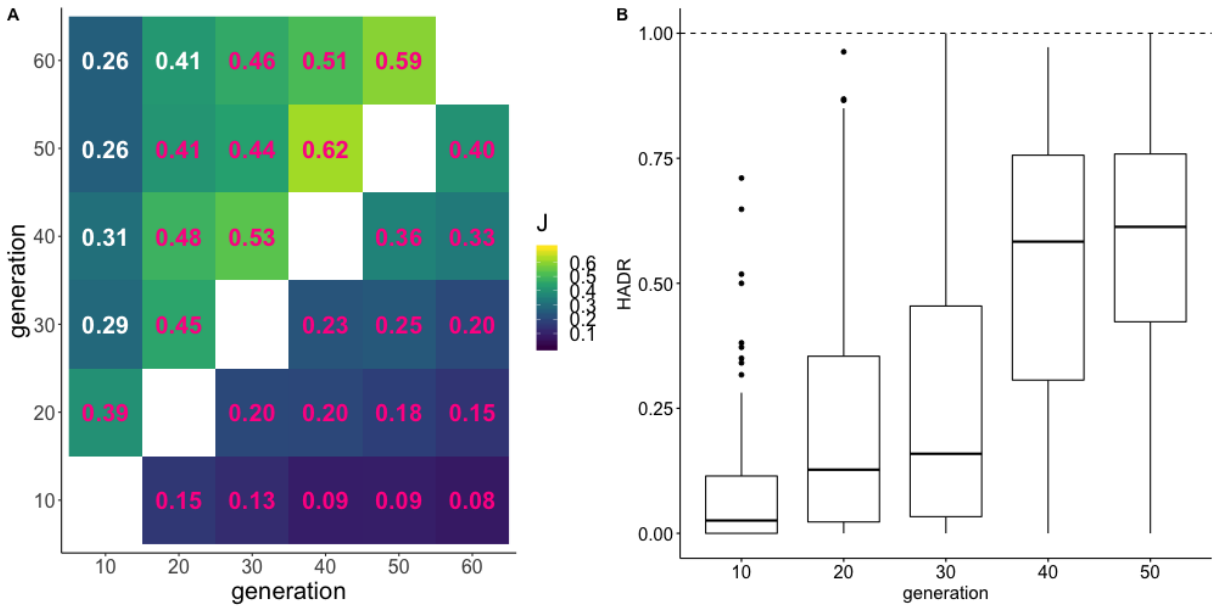

Figure 2: Similarity measures for candidate SNPs, candidate SNPs in a window with a fixed number of SNPs, and candidate SNPs shared with reconstructed selected haplotype blocks
(**A**): Jaccard index (J) for pairwise comparisons of candidate sets. The top triangle shows candidate window sets, the bottom triangle candidate SNP sets. Significant similarities (p-value <0.05 after multiple testing correction, 10 000 bootstraps) are marked in pink.
(**B**): The rate at which selected SNPs of 93 haplotype blocks from generation 60 were already discovered at earlier generations (haplotype discovery rate, HADR).

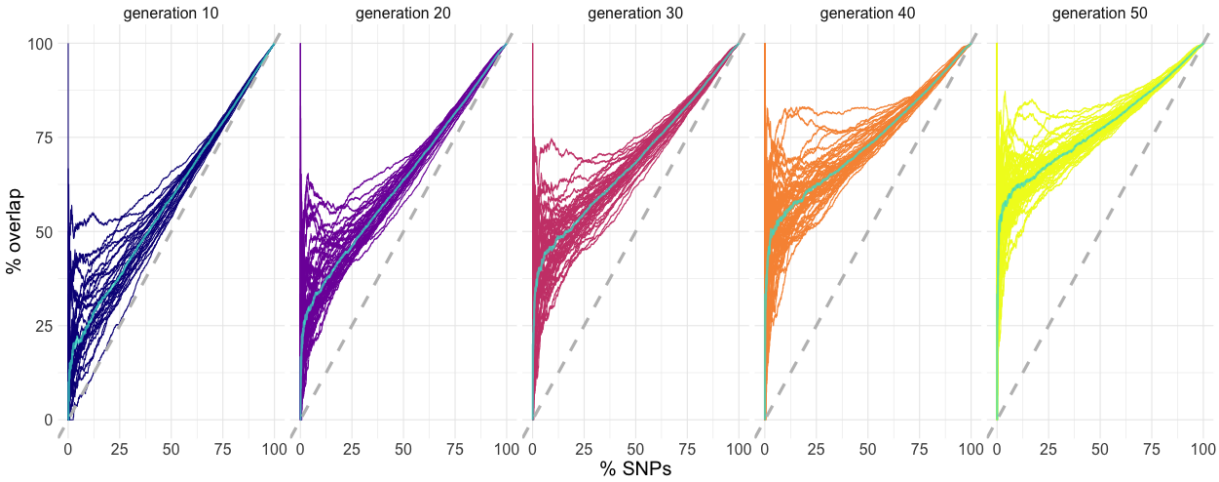

Figure 3: The rank of candidate SNPs becomes more congruent with time. In this ROC-like graph, the ranking of all candidate SNPs in candidate windows is compared. Each panel shows one intermediate time point compared to generation 60. The overlap (in percent) for each candidate window is indicated by a separate line. The median overlap (turquoise line) monotonically increases with experimental duration, demonstrating that the ranking of candidate SNPs is more robust for advanced generations. The grey, dashed lines shows the expected overlap in SNP ranking if every variant at generation 60 is recapitulated in a previous time point.

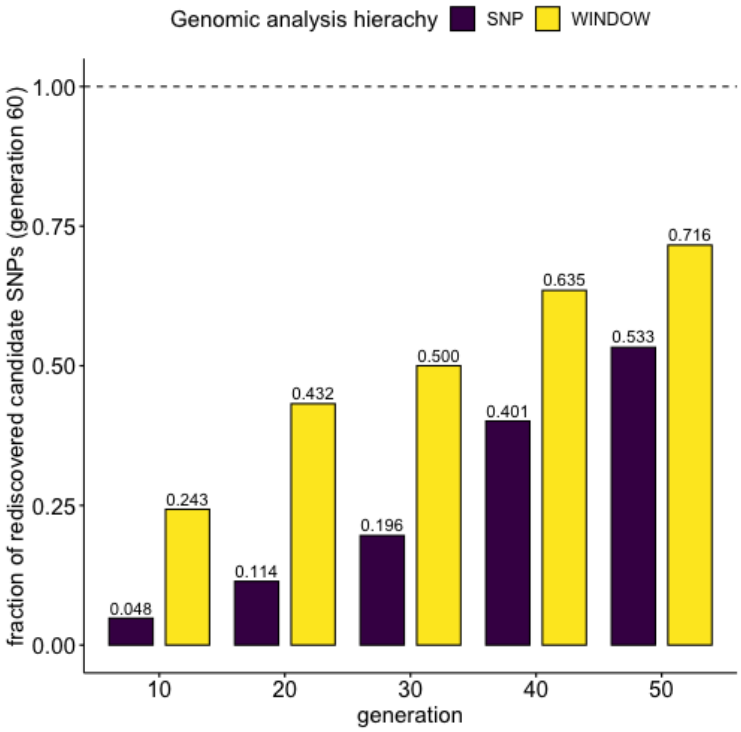

Figure 4: Less than 5% of candidate SNPs in generation 60 are detected consistently at every generation. The bars depict the fraction candidate SNPs (purple) and candidate windows (yellow) at generation 60, which are candidates in all subsequent generations (e.g. 40.1% of generation 60 candidate SNPs are candidates in generation 50 and 40). Candidate windows are more consistent than candidate SNPs. Figure S8 depicts the ratios for candidate sets that are not restricted to SNPs segregating in all generations and time points.

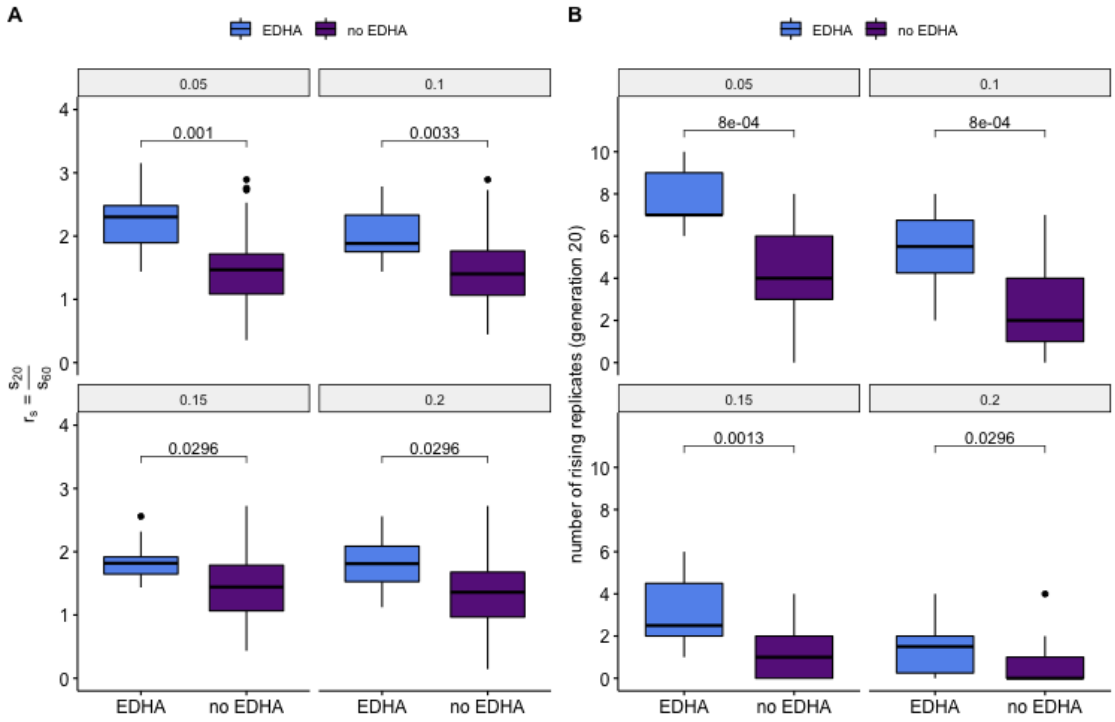

Figure 5: Early Detectable HAplotype blocks (EDHAs) differ from the other selected haplotype blocks. (**A**): The ratio of selection (s_r) coefficients determined for early generations (generation 20) and late generations (generation 60) is significantly higher for EDHAs. (**B**): EDHAs rise in more replicates than other haplotype blocks. Both observations are robust to different allele frequency change thresholds. Values above the boxplots represent the two-tailed Mann-Whitney test *p*-values corrected for multiple testing with the Benjamini-Hochberg procedure.

SUPPLEMENT

List of Figures

1 Genomic analysis hierachy 19
2 Similarity measures for candidate SNPs 20
3 Rank of candidate SNPs over time 21
4 Fraction of shared candidates 22
5 Eearly detectable haplotype blocks properties 23
S1 Time resolved candidate window respresentation s-3
S2 Haplotype discovery rate s-4
S3 Haplotype discovery rate, PCA s-5
S4 Similarity measures for candidate SNPs (including fixed SNPs) s-6
S5 Rank of candidate SNPs over time (simulated data) s-7
S6 Similarity measures for candidate SNPs (simulated data) s-8
S7 Number of detected causative SNPs (simulated data) s-9
S8 Fraction of shared candidates (including fixed SNPs) s-10
S9 Fraction of shared candidates (simulated data) s-11
S10 Early detectable haplotype blocks properties II s-12

S11 Early detectable haplotype blocks properties (simulated data) s-13

List of Tables

S1 N_e estimates s-14
S2 Replicate-specific autosomal N_e estimates s-15
S3 Replicate-specific X N_e estimates s-16
S4 Fraction of shared candidate windows s-17
S5 Number of candidate windows at each time point s-18
S6 Number of candidate SNPs at each time point s-19
S7 Number of candidate SNPs at each time point (simulated data) s-20
S8 Marginal significant candidate SNPs in blocks with fluctuating HADR s-21

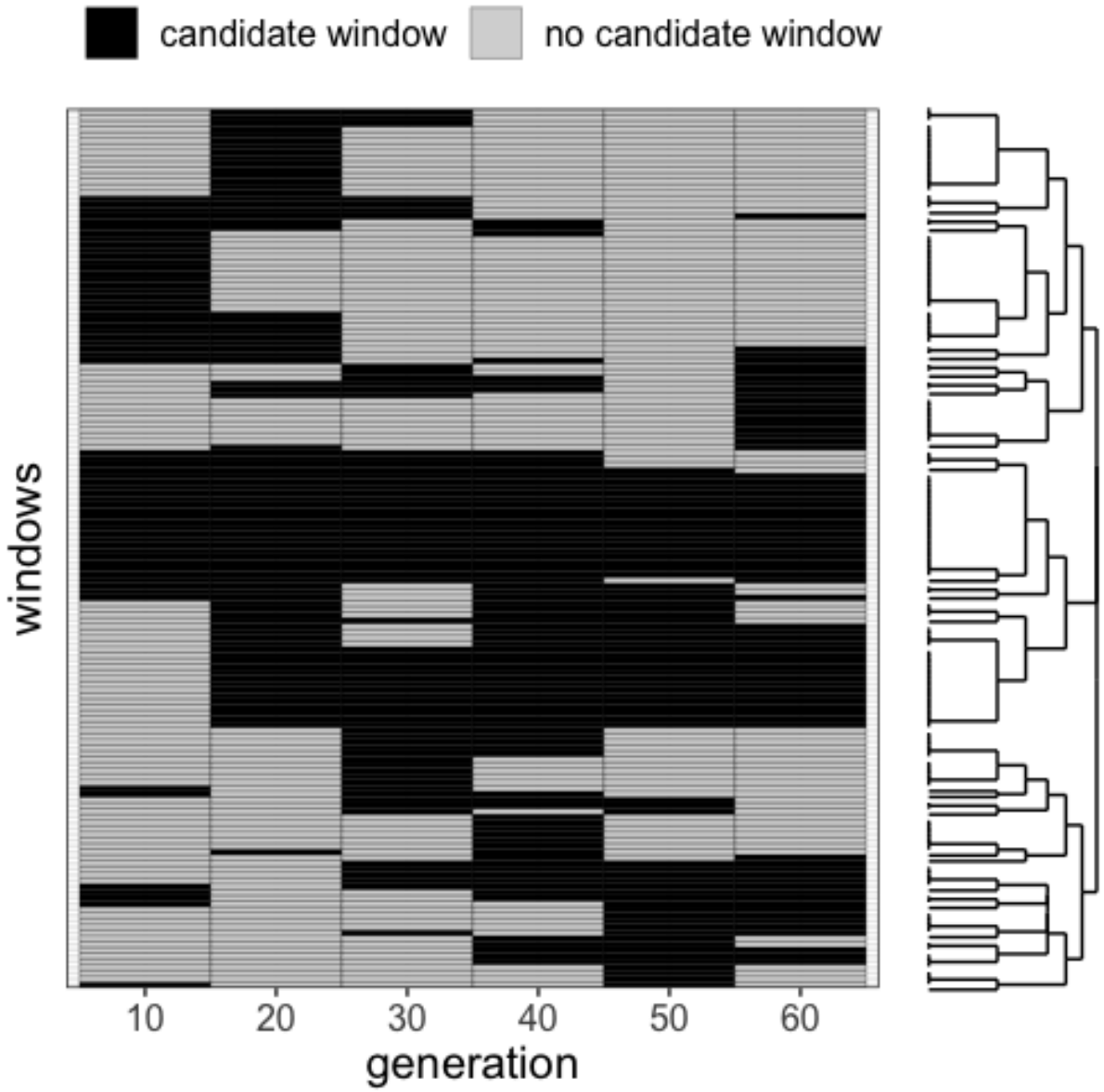

Figure S1: Time resolved representation of all windows that were classified as candidate window for at least one time point: each window is shown in a single row for generation 10 to 60. Windows that are enriched for candidate SNPs at a particular time point are colored black; not enriched windows are colored grey. Windows with similar candidate window patterns are grouped together.

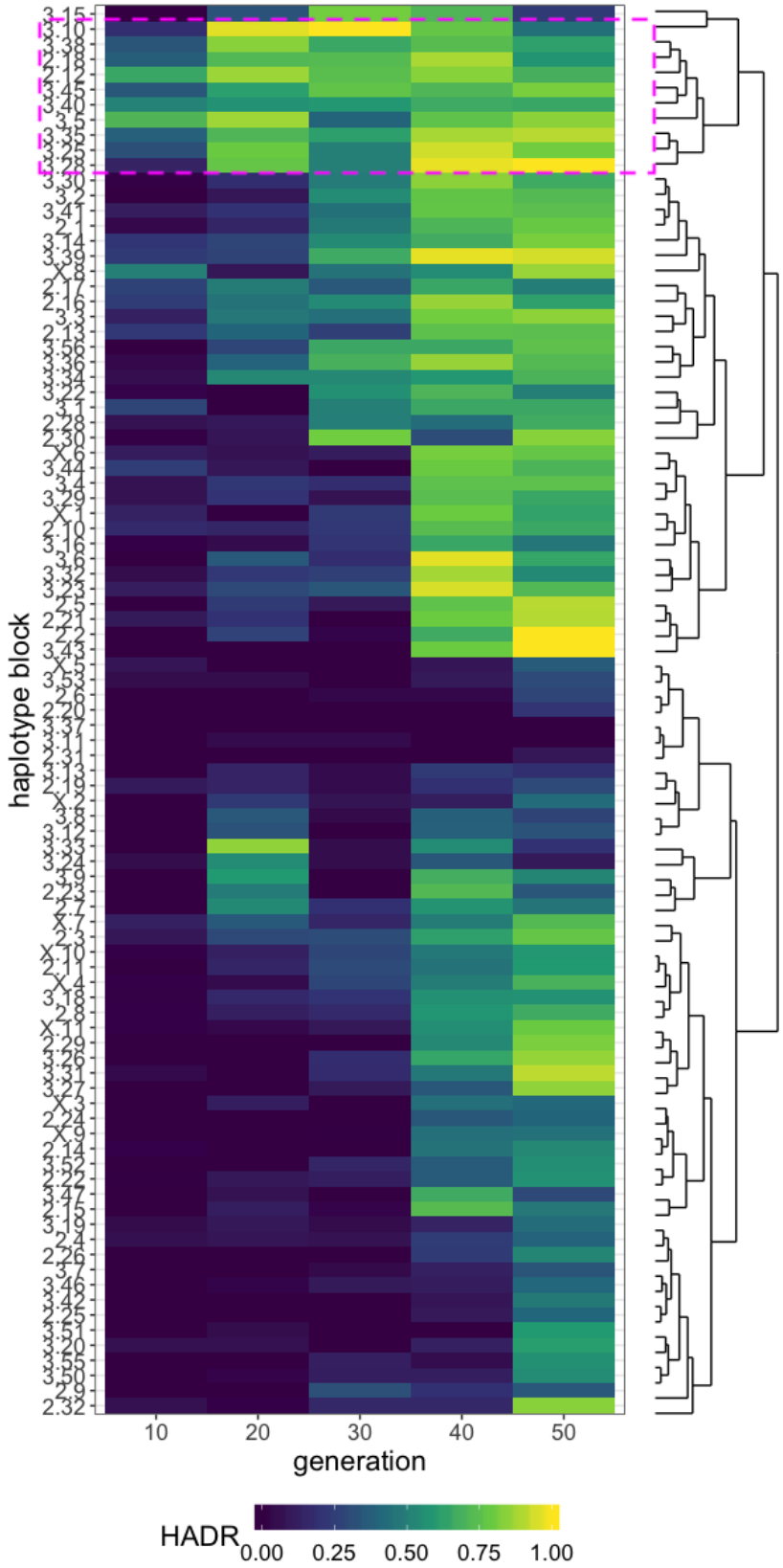

Figure S2: Empirical data: Haplotype discovery rate (HADR) from generation 10 to generation 50. The x-axis shows the five intermediate time points, the y-axis the haplotype blocks determined by Barghi *et al.* (2019). Haplotype blocks are grouped by similar HADR dynamics over time. Tiles are colored by HADR score. The 10 haplotype blocks of the Early Detected Haplotypes (EDHA) cluster (n=10) are highlighted by a pink frame.

While the HADR of most haplotype blocks increases with time, for nine haplotype blocks the HADR score fluctuated (2.7, 2.15, 2.23, 3.8, 3.9, 3.12, 3.24, 3.33, and 3.47). Because these haplotype blocks are mainly based on marginally significant candidate SNPs (Table S8), they are more sensitive to random fluctuations, partly due to experimental noise, partly due to stochasticity of the simulations used to determine the significance threshold.

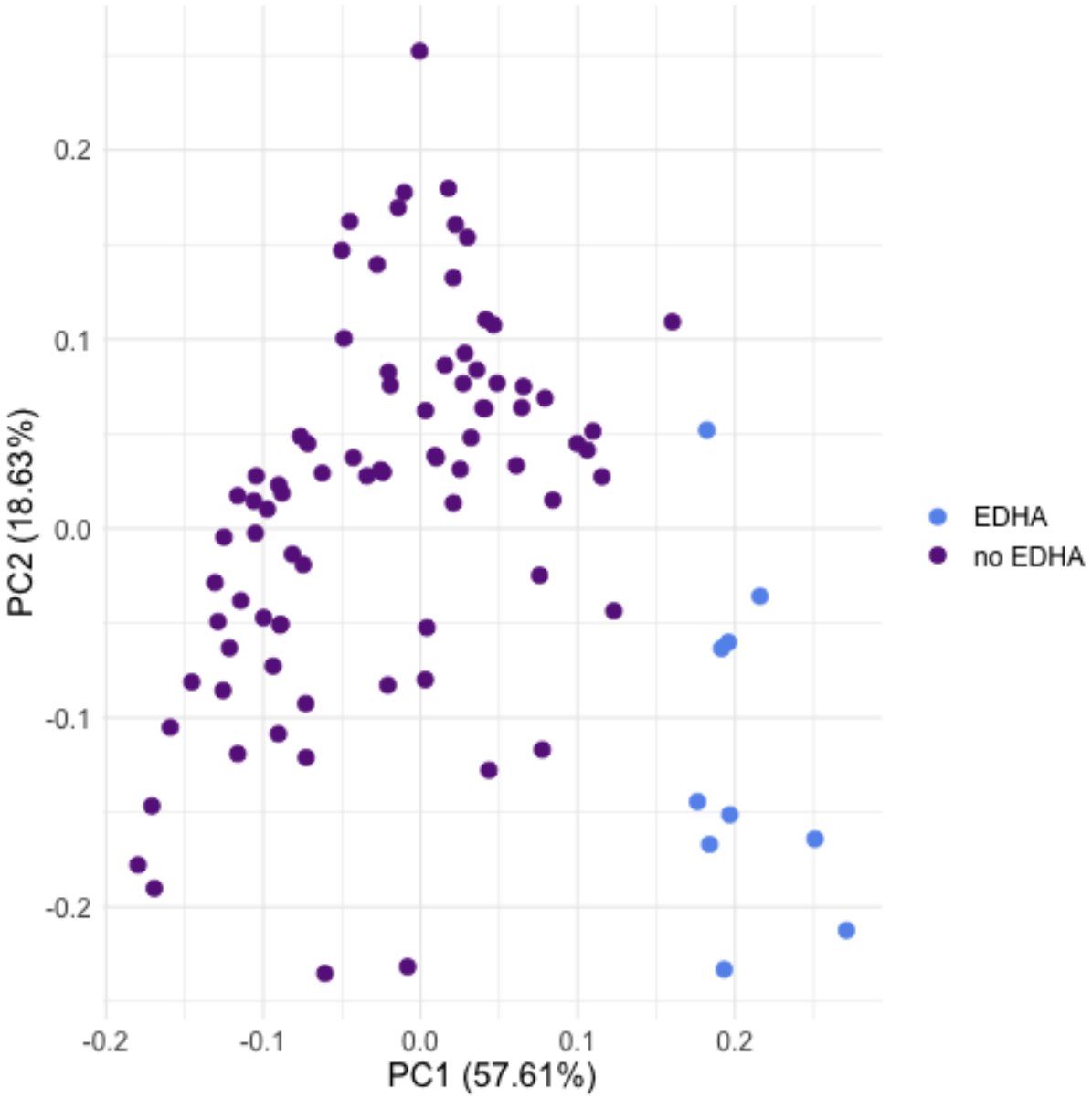

Figure S3: The first two principle components from a PCA based on the haplotype discovery rate (HADR) from generation 10 to generation 50. Each haplotype block is represented by a single dot. Blocks of the early detectable haplotype block (EDHA) cluster are colored in blue, others in purple.

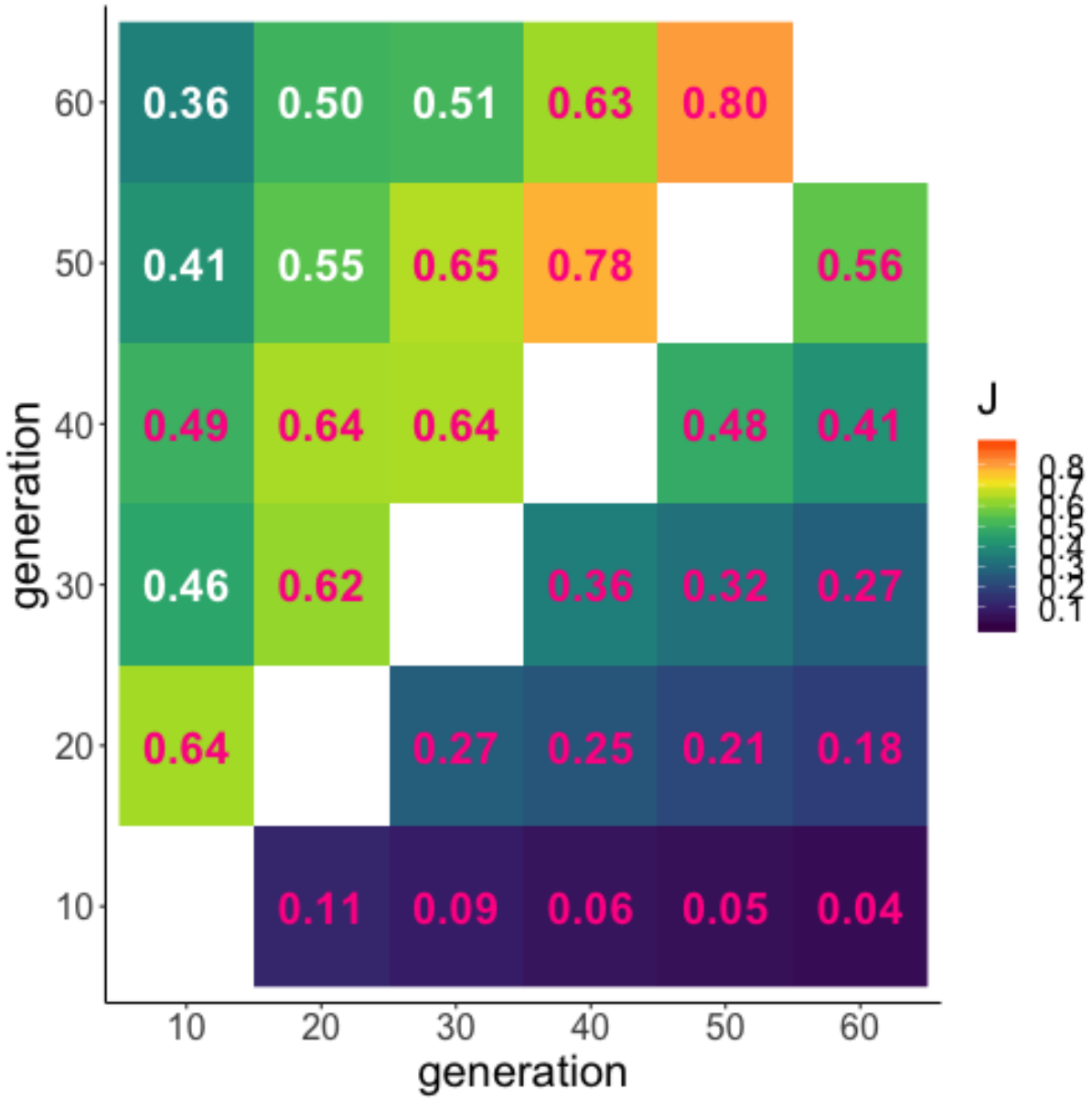

Figure S4: Jaccard index (J) for pairwise comparisons of candidate sets based on a simulated Evolve and Resequence experiment (filtered for SNPs segregating in all generations and time points). The top triangle shows candidate window sets, the bottom triangle candidate SNP sets. Significant similarities (p<0.05 after multiple testing correction, 10 000 bootstraps) are marked in pink.

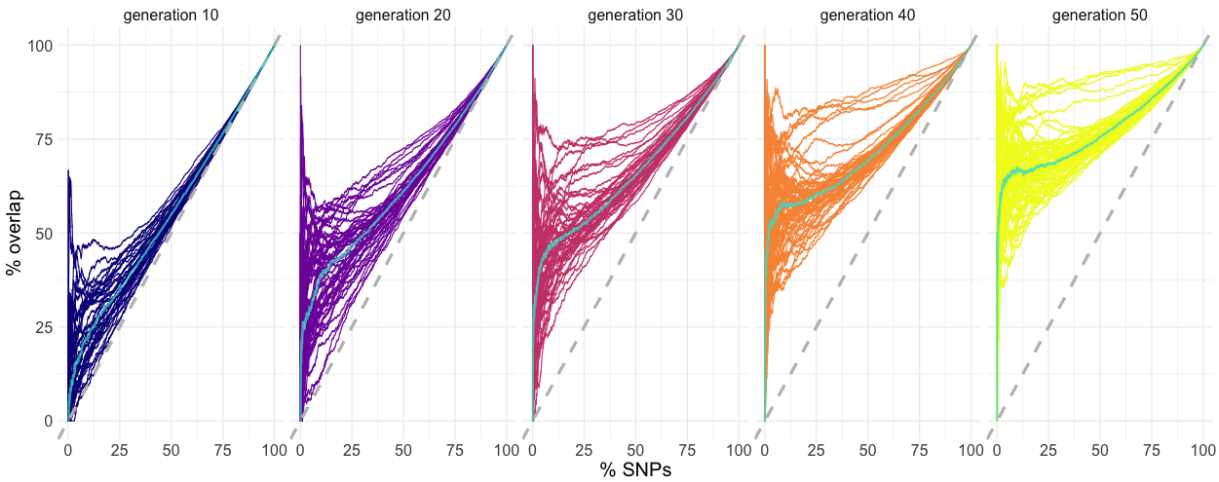

Figure S5: The rank of candidate SNPs becomes more congruent with time in the simulated data. In this ROC-like graph, the ranking of all candidate SNPs in candidate windows is compared. Each panel shows one intermediate time point compared to generation 60. The overlap (in percent) for each candidate window is indicated by a separate line. The median overlap (turquoise line) monotonically increases with experimental duration, demonstrating that the ranking of candidate SNPs is more robust for advanced generations. The grey, dashed lines shows the expected overlap in SNP ranking if every variant at generation 60 is recapitulated in a previous time point.

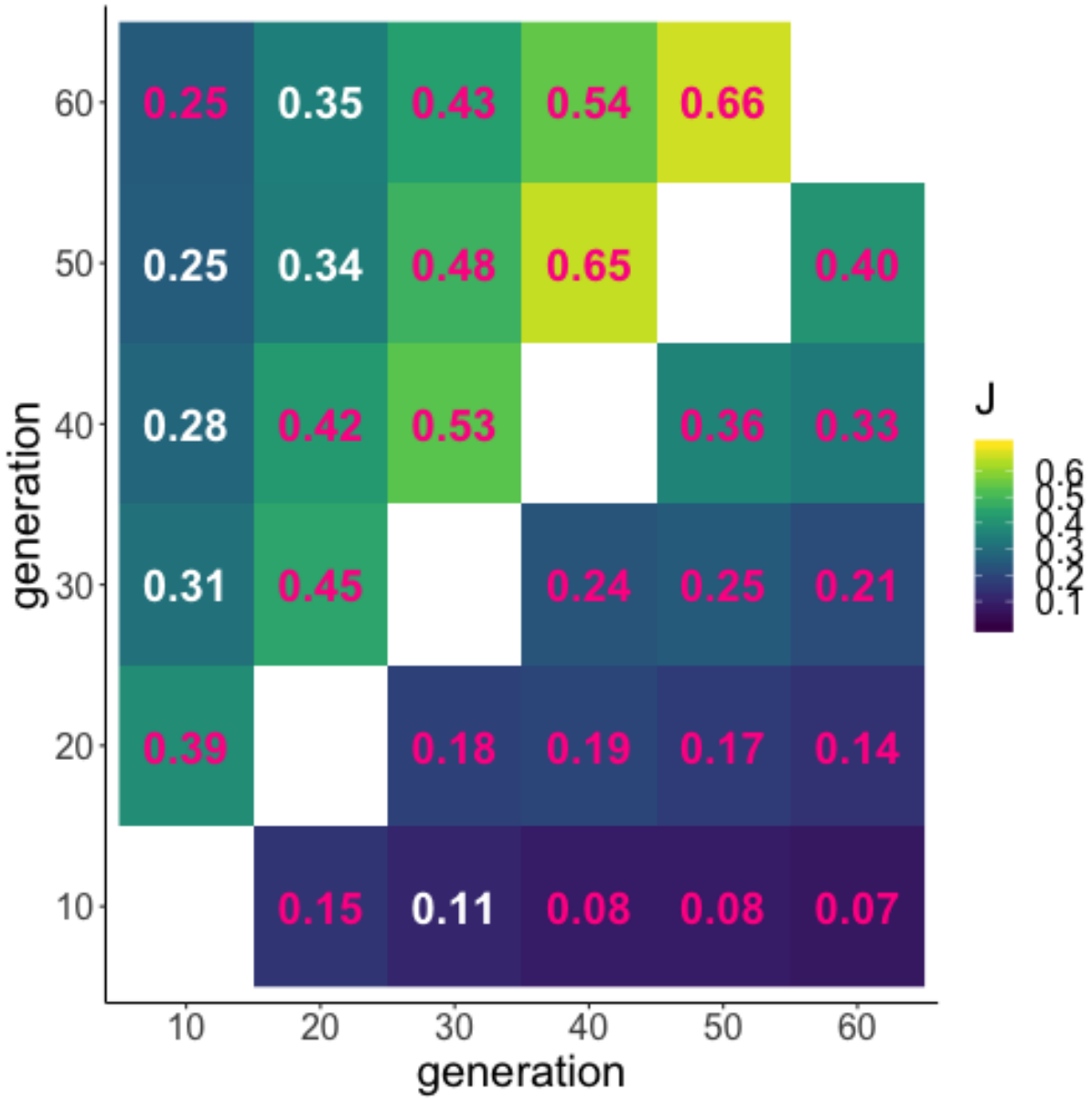

Figure S6: Jaccard index (J) for pairwise comparisons of empirical candidate sets (not filtered for SNPs segregating in all generations and time points). The top triangle shows candidate window sets, the bottom triangle candidate SNP sets. Significant similarities (p<0.05 after multiple testing correction, 10 000 bootstraps) are marked in pink.

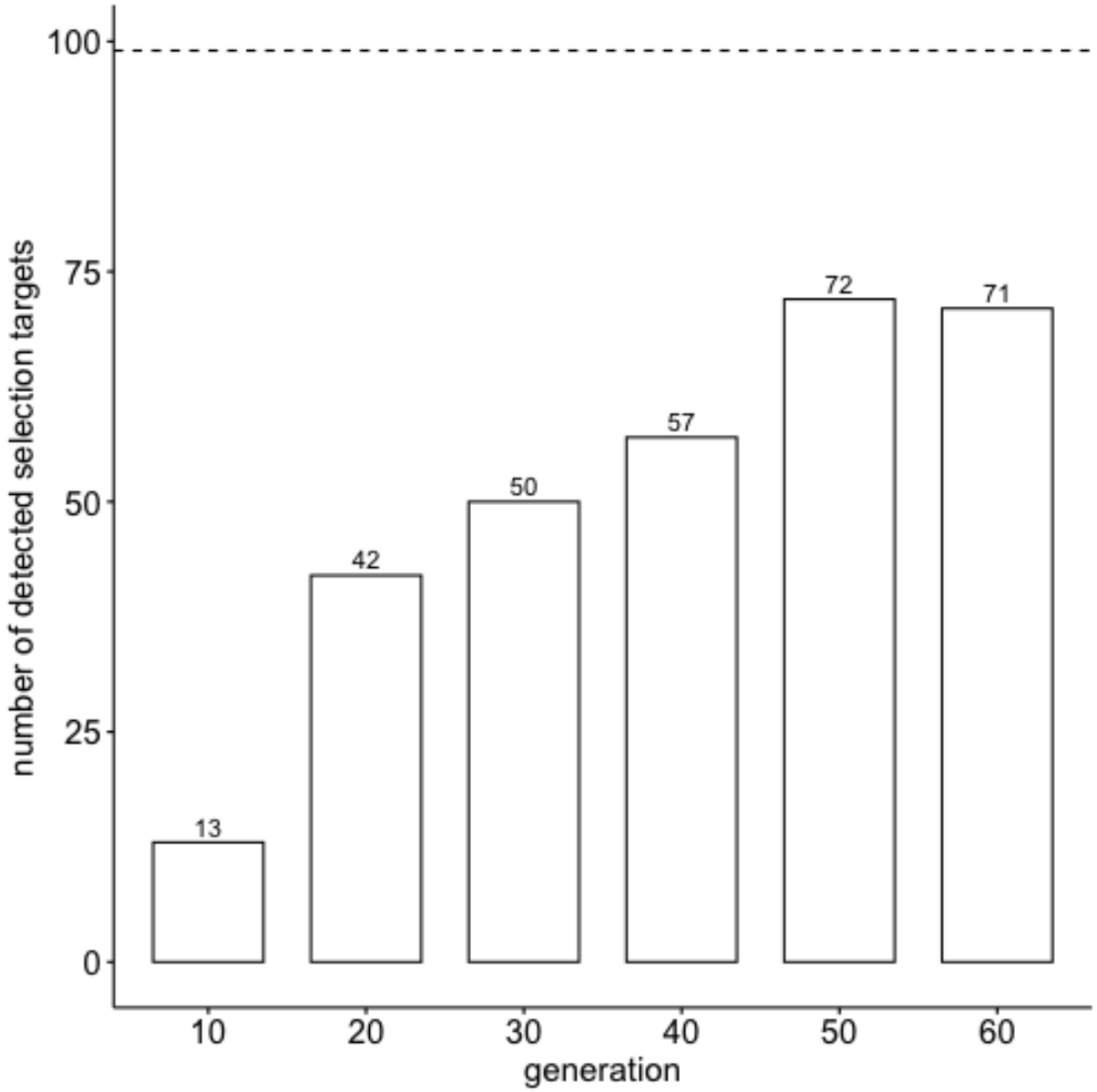

Figure S7: Number of detected causative SNPs for generation 10 to 60. 99 causative SNPs were simulated.

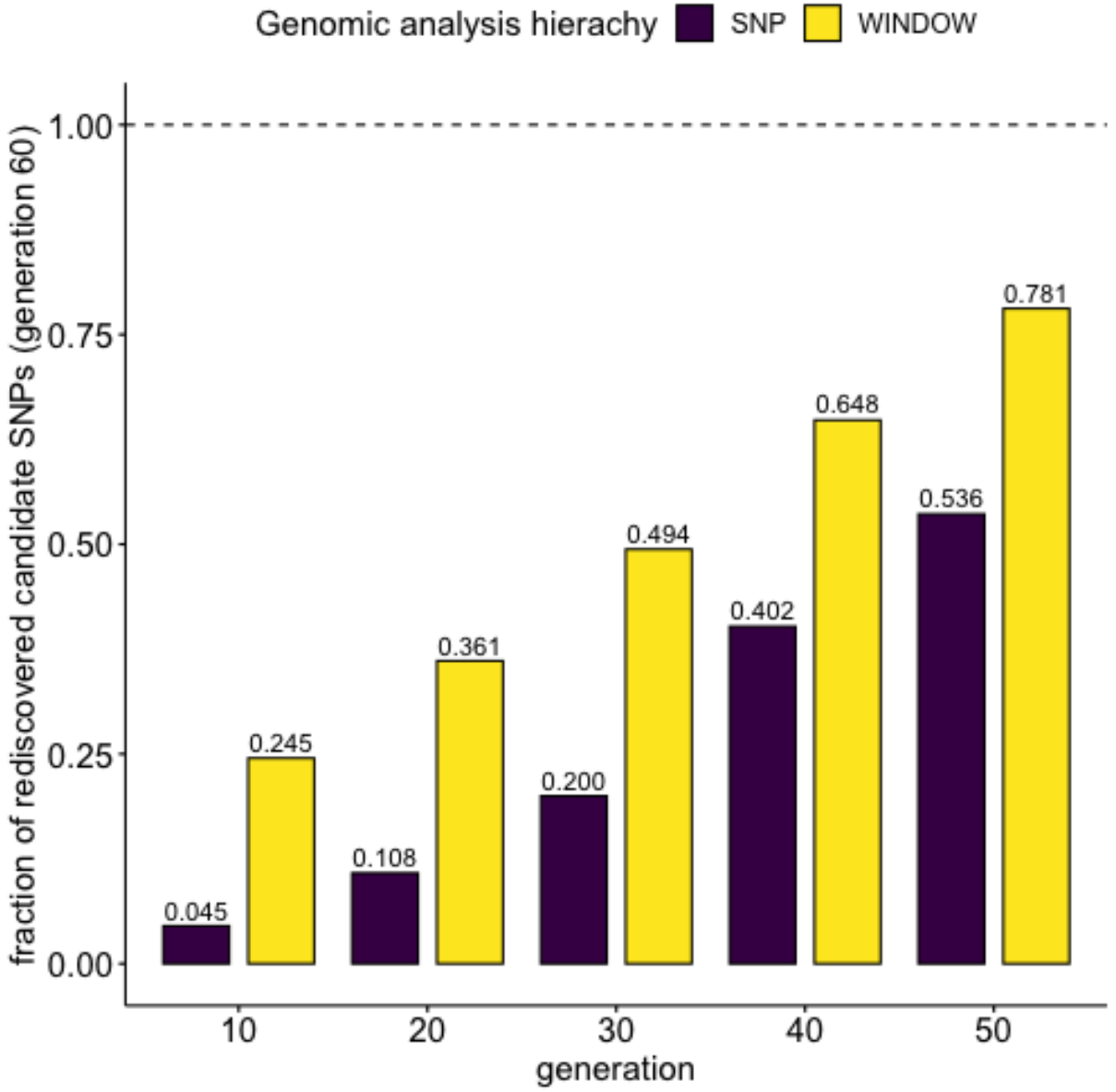

Figure S8: Less than 5% of candidate SNPs in generation 60 are detected consistently at every generation. The bars depict the ratio of candidates at generation 60 (SNPs in purple, windows in yellow) that consistently remain candidates at consecutive earlier time points (e.g. 40.2 % ofthe candidate SNPs at generation 60 are also candidates in generation 50 and 40). Window-based analyses are more consistent than SNP-based analyses.

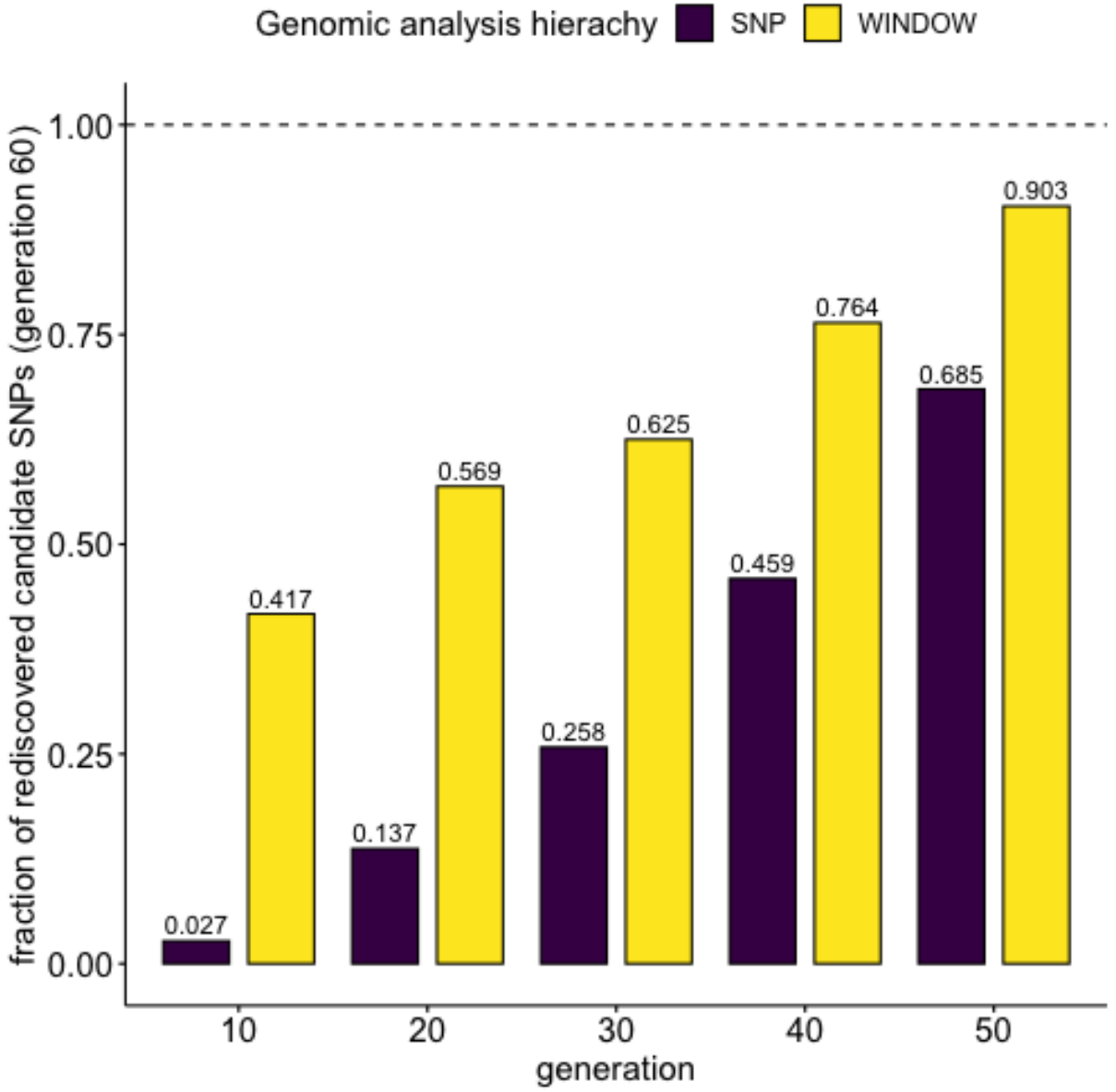

Figure S9: Less than 3% of candidate SNPs in generation 60 are detected consistently at every generation in our simulated Evolve and Resequence experiment. The bars depict the ratio of candidates at generation 60 (SNPs in purple, windows in yellow) that consistently remain candidates at consecutive earlier time points (e.g. 45.9 % of generation 60 candidate SNPs are also candidates in generation 50 and 40). Window-based analyses are more consistent than SNP-based analyses.

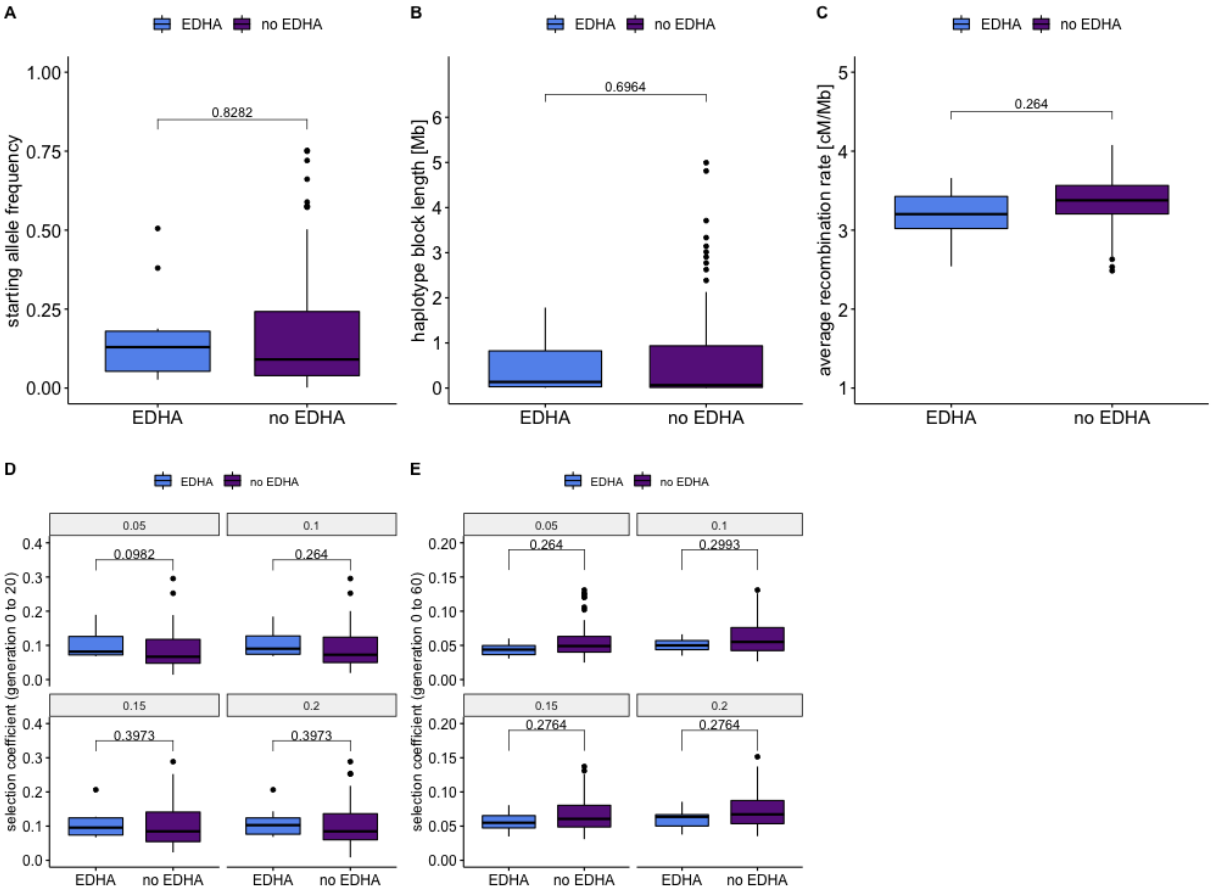

Figure S10: Early detectable haplotype blocks (EDHAs) do not differ in starting allele frequency (**A**), block length (**B**), average recombination rate (**C**), or strength of selection (both early (**D**) and late (**E**) in the experiment) from other haplotype blocks. Values above the boxplots represent the two-tailed Mann-Whitney test *p*-values corrected for multiple testing with the Benjamini-Hochberg procedure.

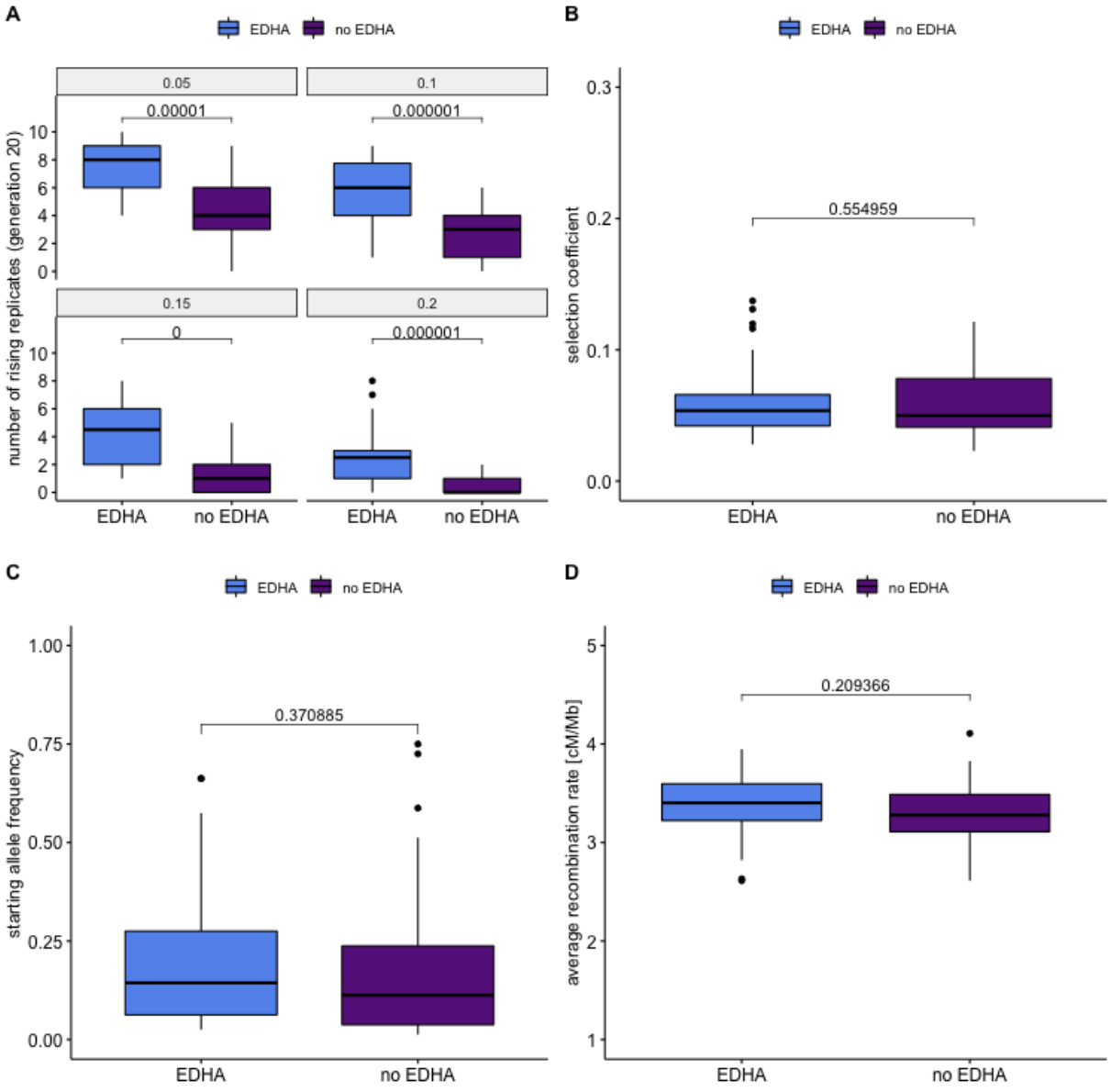

Figure S11: After 20 generations, early detectable targets respond in more replicates than other targets in the simulated Evolve and Resequence experiment – regardless of the chosen allele frequency change cutoff (**A**). Early detectable targets do not differ in selection strength (**B**), starting allele frequency (**C**) or average recombination rate (**D**) from other selected targets. Values above the boxplots represent the two-tailed Mann-Whitney test *p*-values corrected for multiple testing with the Benjamini-Hochberg procedure.

Table S1: N_e estimates for the X-chromosome and autosomes averaged across 10 replicates at different time points

generation	X	autosome
10	257	278
20	230	261
30	237	266
40	239	260
50	245	269
60	262	291

Table S2: Replicate-specific autosomal N_e estimates for different time points

generation	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
10	334	431	333	279	213	251	236	257	262	188
20	309	310	240	208	226	274	221	274	316	235
30	337	304	238	210	234	290	239	243	288	275
40	341	306	238	205	227	287	255	211	270	257
50	356	306	237	235	215	289	281	226	288	257
60	381	336	247	255	246	307	297	244	310	287

Table S3: Replicate-specific X N_e estimates for different time points

generation	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
10	291	411	279	261	209	228	224	251	211	201
20	181	269	220	226	236	191	255	253	240	230
30	190	281	251	207	237	218	233	239	266	253
40	191	268	265	213	220	230	252	217	286	248
50	203	279	260	231	214	236	256	229	292	250
60	217	280	269	240	208	274	270	241	358	268

Table S4: Fraction of candidate windows at generation 60, that are also candidate windows in all previous generations. a: windows size = 5 000 SNPs. SNPs do not necessarily segregate in all replicates and time points. b: window size =13 025 SNPs (a window with 13 025 SNPs result on average in a window length of 250 kb: that is half the step size in the sliding window approach used by Barghi *et al.* (2019) to reconstruct selected haplotype blocks) . SNPs do not necessarily segregate in all replicates and time points. c: window size =5 000 SNPs, which segregate in all replicates and time points.

a	b	c
0.24%	0.28%	0.24%

Table S5: Absolute number and fraction of candidate windows at each generation. a: windows size = 5 000 SNPs. SNPs do not necessarily segregate in all replicates and time points. b: window size =13 025 SNPs. SNPs do not necessarily segregate in all replicates and time points. c: window size =5 000 SNPs, which segregate in all replicates and time points.

generation	a	b	c
10	186 (0.18)	78 (0.20)	62 (0.17)
20	224 (0.22)	109 (0.28)	83 (0.23)
30	196 (0.19)	90 (0.23)	72 (0.20)
40	235 (0.23)	111 (0.28)	83 (0.23)
50	225 (0.22)	92 (0.23)	69 (0.19)
60	233 (0.23)	97 (0.25)	74 (0.21)

Table S6: Number of candidate SNPs at each time point

generation	number of candidate SNPs
10	27 290
20	42 736
30	31 002
40	64 974
50	48 344
60	56 166

Table S7: Number of candidate SNPs at each time point in the simulated Evolve and Resequence data.
true positives: SNPs that are candidate SNPs and s>0
false positives: SNPs that are candidate SNPs and s=0

generation	number of candidates SNPs	true positives	false positives
10	35 954	13	35 941
20	108 129	42	108 087
30	106 441	50	106 391
40	142 696	57	142 639
50	177 052	72	176 980
60	194 573	71	194 502

Table S8: Highest -$\log_{10}$(p−value) in haplotype blocks with HADR fluctuation over time. The first row in the table shows the applied 5 % FDR cut-off for autosomes. The second row contains the highest -$\log_{10}$(p−value) observed. Row three to eleven depict the highest -$\log_{10}$(p−values) for different haplotype blocks. We only depict haplotype blocks with pronounced HADR fluctuations at certain time points.

generation	20	30	40	50	60
cut-off	10	14	15	18	19
max	74.06	66.74	78	102.5	112.85
3.8	15.95	14.1	23.37		
3.12	12.09	13.09	22.02		
3.33	18.65	14.69	19.34		
3.24	22.86	12.48	20.87		
3.9	15.30	13.58	24.56		
2.23	19.17	13.55	24.74		
2.7	17.05	18.76	23.00		
3.47			21.44	26.8	30.85
2.15			32.21	32.63	30.03