Multiple Haplotype Reconstruction from Allele Frequency Data

Supplementary Material

Marta Pelizzola[1,2,a], Merle Behr[3,a], Housen Li[4], Axel Munk[4,5], Andreas Futschik[6, b]

1 Vetmeduni Vienna
2 Vienna Graduate School of Population Genetics
3 University of California Berkeley
4 University of Göttingen
5 Max Planck Institute for Biophysical Chemistry, Göttingen
6 Johannes Kepler University Linz
a These authors contributed equally
b Corresponding author, andreas.futschik@jku.at

Our supplementary material is structured as follows. We first provide additional information on our proposed method in Section S1. In particular, we discuss conditions that ensure identifiability, i.e. unique estimates for our underlying haplotypes and their frequencies. We also provide algorithms and explain how we select the number of haplotypes (model selection), and how accuracy scores are computed that provide information on the quality of the estimates.

In Section S2, we describe our model for the simulations. We provide additional results from our simulations, together with our analysis of the error under several experimental designs, in Section S3. We evaluate the accuracy measure introduced in Section S1-4 with our simulations in Section S4. Furthermore, additional results on the estimation of allele frequencies are provided in Section S5. Section S6 provide an analysis of the simulation runs leading to outliers in the reconstruction error and Section S7 discusses the effects of different levels of recombination on our proposed approach. Lastly, additional results on the real data can be found in Sections S8, S9, and S10.

S1 Theory and Methods

S1-1 Identifiability of structure and frequency from allele frequency (AF)

[Behr and Munk, 2017] derived sufficient and necessary conditions under which the matrices S and W (including the number of haplotypes m) are uniquely identifiable from their product SW. With some slight modifications of their arguments, we can also show that under weak identifiability assumptions on S and W, one can uniquely identify S, W, and b from the population AFs F. More precisely, for W it is assumed that different combinations of SNPs lead to different AFs, that is,

$$sW \neq s'W \text{ for all } s \neq s' \in \{0,1\}^m. \tag{S1}$$

For the haplotype structure S it is assumed that there is at least one SNP which is unique to a haplotype and at least one SNP that is only present in minor haplotypes, that is

$$\begin{aligned} &\text{for all } i \in [m] \text{ there exists an } n \in [N] \text{ such that} \\ &\qquad S_{ni} = 1 \text{ and } S_{nj} = 0 \text{ for all } j \neq i \\ &\text{and there exists an } n \in [N] \text{ such that } S_{ni} = 0 \text{ for all } i \in [m], \end{aligned} \tag{S2}$$

(equivalently one can exchange 0 and 1 in (S2)). Both of these conditions are very reasonable in most real data situations, given that the number of essential haplotypes m is not too large. It is easy to see that condition S1 is necessary for identifiablity of haplotype structure S and frequency W from AF Y in (2). A simple situation, where S1 does not hold is when two haplotypes have exactly the same proportion at all time points $t \in [T]$. In that case, it is not possible to distinguish whether a SNP is present in one or the other haplotype. Condition (S2) imposes a sufficient variability of individual haplotypes. A trivial non-identifiable counter example is, for instance, when one major haplotype is constant zero or constant one. Some further insights and examples on the specific condition in (S2) can be found in [Behr and Munk, 2017, Behr et al., 2018]. Note that (S2) requires that out of the 2^m possible variant combinations for the m haplotypes, at least those m combinations which correspond to the identity vectors $e_1 = (1, 0, \dots, 0), \dots, e_m = (0, \dots, 0, 1)$ and the one which corresponds to the zero vector $(0, \dots, 0)$ appear at some of the locations $n \in [N]$.

The conditions (S1) and (S2) do not just guarantee identifiability in an abstract way, but they also lead to an explicit algorithm for recovering S, W, b and m from the noiseless AFs $SW + b$ in (2). Part of our reconstruction algorithm is built on this deterministic recovery algorithm that is based on a simple combinatorial reordering of the observations (see [Behr and Munk, 2017, Diamantaras and Chassioti, 2000] for very similar algorithms). The idea of this algorithm is that the discrete nature of S lets us identify both S and W from appropriate row vectors of Y as outlined in the following.

The smallest norm among the rows of Y appears for any SNP that has variant 0 for all m haplotypes, in which case we observe only the bias term b. Similar, the second (and third) smallest possible row value of Y appears for a SNP with variant 0 on all haplotypes, except the one with the smallest frequency $W_{m\cdot}$ (second smallest frequency $W_{(m-1)\cdot}$), which lets us identify $W_{m\cdot}$ and $W_{(m-1)\cdot}$. Among the remaining observed row values of Y the smallest one must correspond to $W_{(m-2)\cdot}$, and so on. In that way, one can successively recover all the frequencies $W_{i\cdot}$ and given W it is straightforward to recover S. We present pseudo code in Algorithm 1 below.

S1-2 Algorithms

Algorithm 1 Recover S, W, b from exact data $Y = SW + b$

procedure HAPLOSEPCOMBIEXACT
Input: $Y = SW + \mathbf{1}b^\top$ such that (S1) and (S2) hold.
Output: S, W, b, m
  $\mathcal{Y} \leftarrow \{Y_{1\cdot}, \ldots, Y_{N\cdot}\}$
  $b \leftarrow \arg\min_{y \in \mathcal{Y}} \|y\|$
  $\mathcal{Y} \leftarrow \mathcal{Y} \setminus b$
  $\mathcal{Y} \leftarrow \mathcal{Y} - b$
  $W_{1\cdot} \leftarrow \arg\min_{y \in \mathcal{Y}} \|y\|$
  $\mathcal{Y} = \mathcal{Y} \setminus W_{1\cdot}$
  $m \leftarrow 1$
  while $\mathcal{Y} \neq \emptyset$ **do**
    $W_{(m+1)\cdot} \leftarrow \arg\min_{y \in \mathcal{Y}} \|y\|$
    $m \leftarrow m + 1$
    $\mathcal{Y} \leftarrow \mathcal{Y} \setminus \{\sum_{i=1}^{m} s_i W_{i\cdot} \ : \ s \in \{0,1\}^m\}$
  end while
  for $n = 1$ to N **do**
    $S_{ni} \leftarrow \arg\min_{s \in \{0,1\}^m} \|Y_{n\cdot} - sW\|$
  end for
  put $W_{i\cdot}$ in the reverse order
  return S, W, b, m
end procedure

Algorithm 2 Recover S, W, b from Y in (2)

procedure HAPLOSEP
Input: $Y \in [0,1]^{N\times T}, m \in [N], \delta > 0$
Output: $\hat{W}, \hat{b}, \hat{S}$
  $(\hat{W}, \hat{b}) \leftarrow$ HAPLOSEPCOMBI(Y, m)
  for $n = 1$ to N **do**
    $\hat{S}_{ni} \leftarrow \arg\min_{s\in\{0,1\}^m} \|Y_{n\cdot} - s\hat{W} - \mathbf{1}\hat{b}^\top\|$
  end for
  $E_0 \leftarrow 0$
  $E_n \leftarrow \|Y - \hat{S}\hat{W} - \mathbf{1}\hat{b}^\top\|$
  while $|E_n - E_0| > \delta$ **do**
    $E_0 \leftarrow E_n$
    $(\hat{W}, \hat{b}) \leftarrow \arg\min_{W,b} \|Y - \hat{S}W - \mathbf{1}b^\top\|$
    such that $W_{it}, b_t \in [0,1]$, $\sum_{i=1}^m W_{it} \leq 1$
    for $n = 1$ to N **do**
      $\hat{S}_{ni} \leftarrow \arg\min_{s\in\{0,1\}^m} \|Y_{n\cdot} - s\hat{W} - \mathbf{1}\hat{b}^\top\|$
    end for
    $E_n \leftarrow \|Y - \hat{S}\hat{W} - \mathbf{1}\hat{b}^\top\|$
  end whilereturn $\hat{W}, \hat{b}, \hat{S}$
end procedure

S1-3 Model selection via SVD

Note that in the noiseless population case ($Y = SW + b$ in (2)) the number of dominant haplotypes m can directly be obtained via the rank of the AF matrix with

$$\operatorname{rank}(SW + \mathbf{1}b^\top) = m + 1. \tag{S3}$$

To see this, note that the tth column of $SW + b$ can be written as

$$\sum_{i=1}^{m} S_{\cdot i} W_{it} + b_t (1, \ldots, 1)^\top$$

and thus

$$\operatorname{rank}(SW + \mathbf{1}b^\top) = \dim(\operatorname{span}(S_{\cdot 1}, \ldots, S_{\cdot m}, (1, \ldots, 1)^\top)) = m + 1,$$

where the last equality follows from the identifiability condition (S2). Thus, estimation of m from Y corresponds to estimating the (low) rank of the matrix $SW + b$ from its noisy version Y. A more general strategy for the noisy case is to consider the singular values $s_1, \ldots, s_T$ of Y (assuming that $N \geq T$) and then estimate

$$\hat{m} + 1 = \#\{s_i \geq \tau \ : \ i \in [\min(N, T)]\} \tag{S4}$$

for some threshold τ. [Gavish and Donoho, 2014] derived optimal thresholds (in terms of matrix denoising) that are approximately

$$\tau \approx (0.5(T/N)^3 - 0.95(T/N)^2 + 1.82(T/N) + 1.43)\ s_{\text{med}}, \tag{S5}$$

where s_{med} denotes the median of the singular values $s_1, \ldots, s_T$ of Y. In summary, we estimate $\hat{m}$ as in (S4) with τ as in (S5).

Algorithm 3 Initialize $\hat{W}, \hat{b}$ from Y in (2)

procedure HAPLOSEPCOMBI
Input: $Y \in [0,1]^{N \times T}$ and $m \in [N]$
Output: $\hat{W}, \hat{b}$
  $\{C_1, \ldots, C_{2^m}\} \leftarrow$ apply hierarchical clustering to $\{Y_{n\cdot} : n \in [N]\}$ with 2^m centers $\subset [0,1]^T$.
  $\hat{\mathcal{C}} \leftarrow \{C_1, \ldots, C_{2^m}\}$
  $\hat{b} \leftarrow \arg\min_{c \in \hat{\mathcal{C}}} \|c\|$
  $\hat{\mathcal{C}} \leftarrow \hat{\mathcal{C}} \setminus b$
  $\hat{\mathcal{C}} \leftarrow \hat{\mathcal{C}} - b$
  $\hat{W}_{1\cdot} \leftarrow \arg\min_{c \in \hat{\mathcal{C}}} \|c\|$
  $\hat{\mathcal{C}} \leftarrow \hat{\mathcal{C}} \setminus \hat{W}_{1\cdot}$
  for $l = 2$ to m **do**
    $\hat{W}_{l\cdot} \leftarrow \arg\min_{c \in \hat{\mathcal{C}}} \|c\|$
    for $s \in \{0,1\}^{l-1}$ **do**
      $\hat{\mathcal{C}} \leftarrow \mathcal{C} \setminus \{\arg\min_{c \in \hat{\mathcal{C}}} \|c - \sum_{i=1}^{l-1} s_i \hat{W}_{i\cdot} - \hat{W}_{l\cdot}\|\}$
    end for
  end forreturn $\hat{W}, \hat{b}$
end procedure

S1-4 Accuracy scores

In practice, it may happen that our modeling assumption of a small number of major haplotypes $m \ll T, N$ is violated, e.g., because only few haplotypes are lost over time under some neutral scenario without selection. Alternatively, the selected haplotypes may get lost early on due to random genetic drift. In such a case, a low dimensional haplotype representation will often yield a poor fit to the data Y, which we measure using the well known coefficient of determination $R^2 = 1 - \frac{\|Y - \hat{S}\hat{W} - \mathbf{1}\hat{b}^\top\|^2}{\|Y - \overline{Y}\|^2}$. Besides R^2, we also report the uncertainty of the proposed estimates via bootstrap confidence scores and bands [Efron, 1979]. Recall that the haplotype structure S is constant over the time points $t \in [T]$. Thus, in order to evaluate uncertainty in the estimate $\hat{S}$, we propose to resample (with replacement) from the empirical distribution on $\{Y_{\cdot 1}, \ldots, Y_{\cdot T}\}$, that is,

$$Y_t^\star \overset{\text{i.i.d.}}{\sim} \frac{1}{T} \sum_{t=1}^{T} 1_{Y_{\cdot t}}, \tag{S6}$$

where 1_y denotes the dirac measure on y. For each haplotype $i \in [m]$ and SNP location $n \in [N]$ via sampling $Y^\star = (Y_1^\star, \ldots, Y_T^\star)$ from (S6), we compute the variance of $\hat{S}_{ni}(Y^\star)$. As stability score for the ith haplotype estimate we report the following score:

$$\text{StabScoreS}_i = 1 - \frac{1}{N} \sum_{n=1}^{N} |\hat{S}_{ni} - \frac{1}{K} \sum_{k=1}^{K} \hat{S}_{kni}| \in [0,1]. \tag{S7}$$

A stability score of $\text{StabScoreS}_i = 1$ suggestes an unbiased estimate of the ith haplotype and stability score of $\text{StabScoreS}_i = 0$ a highly biased estimate, which may occur due to model misspecification (i.e., violation of the major haplotype assumption or the identifiability conditons).

For the haplotype frequencies W, we observe that they are invariant for different locations

$n \in [N]$. Thus, to evaluate uncertainty for W we resample from

$$Y_n^* \overset{\text{i.i.d.}}{\sim} \frac{1}{N} \sum_{n=1}^{N} 1_{Y_{n\cdot}} \tag{S8}$$

We report the 0.025 and 0.975 quantiles of $\hat{W}_{it}(Y^*)$ as bootstrap confidence bands and the average width of those confidence bands as stability scores.

In practice, we found the above scores to perform reasonable, but we clearly note that there are many other possibilities to construct quality scores for our setting, such as other bootstrap based scores, or also Bayesian credible scores, or frequentist p-values that are based on explicit modeling assumptions, potentially conditioning on either $\hat{W}$ or $\hat{S}$ to construct conditional confidence statements for the other.

We determine a criterion for accepting scenarios where the reconstruction has enough accuracy overall and consider the structure and frequency specific accuracy scores only for those scenarios. Our criterion is based on the R^2 scores and the frequency change of the haplotype reaching highest frequency. More specifically, we require $R^2 > 0.8$ and the frequency change of the haplotype reaching highest frequency > 0.1.

S2 Simulation setup

We evaluate our approach using extensive simulations. In our simulations we considered three experimental designs aiming to reproduce the three data sets we analyze in Section 5, i.e. the experiments explained in [Noble et al., 2019], [Castro et al., 2019] and [Barghi et al., 2019]. They cover three very different organisms used in E&R experiments (*Caenorhabditis elegans*, mice, and *Drosophila simulans*) with various complexities leading to three different starting conditions for the experiments. Indeed, mice populations need to be small because of the maintenance effort involved, whereas this is not the case for *Drosophila simulans* and even less for *Caenorhabditis elegans*. The latter two organisms thus give more freedom to choose the number of different starting haplotypes.

Selection is an important factor in E&R experiments where researchers attempt to understand the genetic architecture of adaptation. In the literature, several E&R experiments have been discussed that involve different stressful conditions. Sources of stress can be high/low-quality food, body size constraints (e.g. only sufficiently small or large organisms are allowed to reproduce), or heat. Our three data sets consider stress conditions on the reproduction regime [Noble et al., 2019], on the body size [Castro et al., 2019] and the temperature regime [Barghi et al., 2019]. Other publications focus on desiccation resistance [Griffin et al., 2017], pathogen resistance [Kraaijeveld and Godfray, 2008], and selection on flying speed [Weber, 1996].

In our simulations, we consider starting populations with the same numbers of haplotypes, and of individuals, as in the real data applications discussed in Section 5. As some of the founder haplotypes from [Barghi et al., 2019] were made available to us by the authors, starting populations were obtained by sampling from these haplotypes. For our basic scenario, we introduce a simple selection regime with selection strength $s = 0.05$ for a beneficial allele present at three different founder haplotypes. The genetic composition of generation n is obtained by multinomial sampling from the previous generation. Sequencing data are generated every tenth generation at 16 different time points $(G_0, G_{10}, \ldots, G_{150})$. From the simulated haplotype data, we compute the true allele frequencies via the regression model $Y = SW$ in Section 3 of the main text as the matrix product of the simulated haplotype structure and frequency. Afterward, we simulate observed allele frequencies using binomial sampling with sample size n equal to the local sequencing coverage, taken from a Poisson(80) distribution. This is to mimic that real allele frequency data in most E&R experiments are noisy because individuals are sequenced as

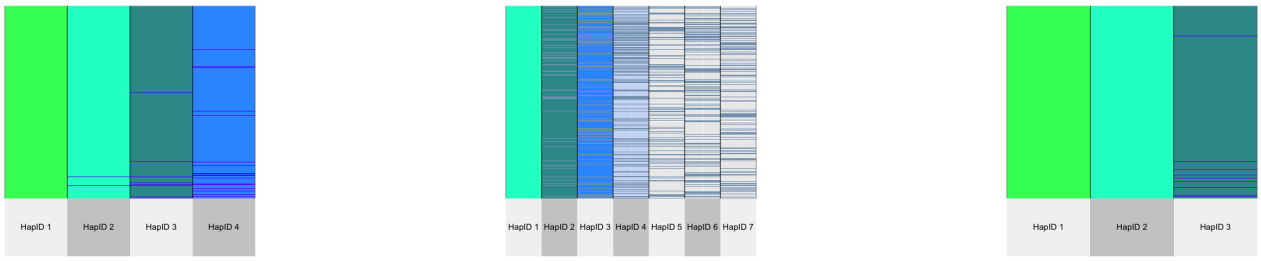
a pool with a given depth (coverage) that changes according to the available resources. With pool sequencing the DNA of all organisms is mixed and sequenced together. An extensive explanation of pool sequencing can be found in [Schlötterer et al., 2014]. A detailed description of this binomial sampling step can be found in [Waples, 1989] and [Jónás et al., 2016].

Our haplotypes involve a genomic region of 500 linked SNPs. We chose such a fairly short window size, because it makes recombination events sufficiently rare to be ignored for our organisms during the experiment. To extend the haplotype reconstruction to larger genomic regions, we propose to apply our method to overlapping sliding windows.

Beyond our basic scenario, we also investigate several alternative scenarios, and consider how design parameters of E&R experiments affect the quality of our haplotype reconstruction. Parameter values not mentioned in our results have been chosen as in our basic scenario.

S3 Simulation results

Complementing Section 4.1, we provide results for our three simple selection scenarios on the comparison between the reconstructed and the true haplotype structure in Fig. S1.



S1a
Longshank mice exp.

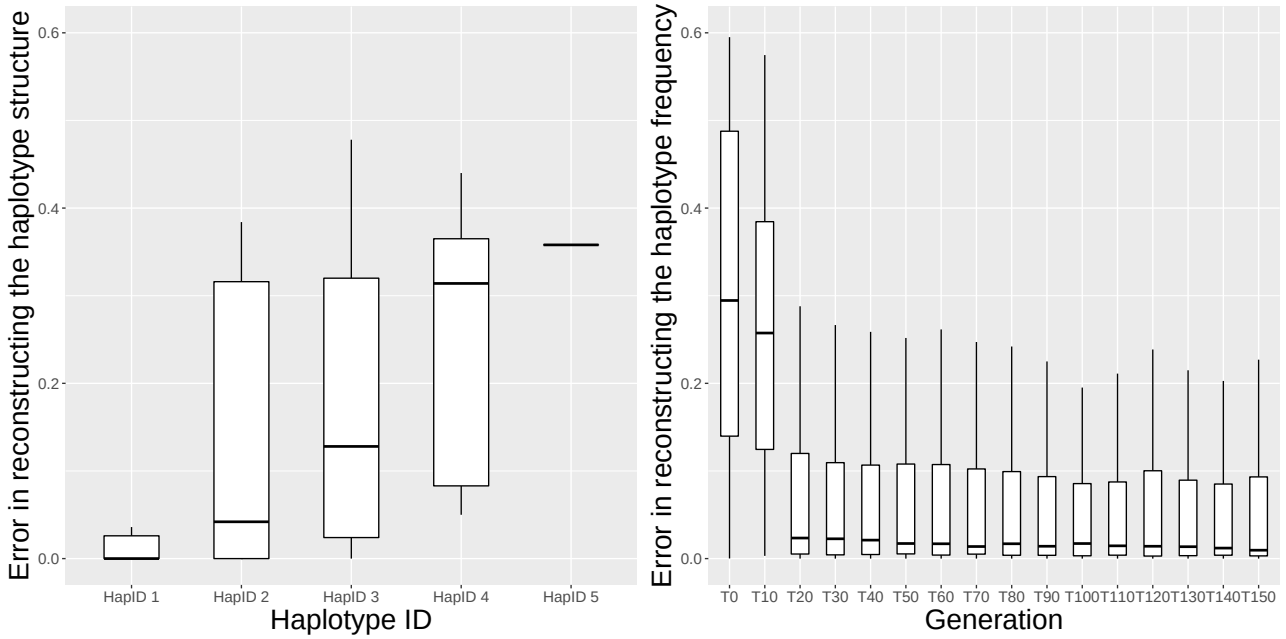
S1b *C. elegans*

S1c *D. simulans*

Figure S1: Result of one simulation run from the simple selection scenario with the experimental design from the Longshank mice experiment (a), *C. elegans* (b), and *Drosophila simulans* (c). This figure shows inconsistencies between true and reconstructed haplotype structure. Blue line indicates mismatches.

Most of the mismatches that we observe in Fig. S1 are in the low-frequency haplotypes. In order to reconstruct haplotypes correctly, they need to be present in the population at an appreciable frequency for several generations. In particular our approach usually cannot accurately reconstruct the structure of haplotypes reaching zero frequency in the earlier part of the experiment. Even so, those haplotypes are not of interest for most analyses trying to understand the architecture of adaptation because they do not provide any contribution to it. Since the number of true haplotypes can be much larger than the number of haplotypes we reconstruct, we match the (true) haplotype having the closest possible structure to the given reconstructed one to compute the error for our estimated haplotypes. As for the figures in the main text, we filter again using our criteria on R^2 and the frequency change of the most abundant haplotype as explained in Section S1-4. See Section S6 for the remaining simulation runs. Based on 100 simulation runs, Fig. S2 shows very low error for both frequency and structure of the selected haplotype(s). However, looking at the different time points, the error is higher for initial generations, whereas it drops for later stages of evolution (see Fig. S2b). The differences between earlier and later time points can be pronounced depending on the experimental design. Indeed when selection occurs, our method provides better estimates for later time points than for earlier ones, if the number of reconstructed haplotypes is much smaller than the number of haplotypes in the starting population. Similar conclusions can be drawn also for the results about the experimental design based on [Noble et al., 2019], shown in Fig. S3.

Starting from these three simple selection scenarios, we did simulations for different values of important parameters for E&R in order to assess how they affect our haplotype reconstruction. We focus on the selection coefficient, the number of haplotypes in the founder population, the number of haplotypes carrying the beneficial allele, the coverage and the number of time points where the sequencing data are collected. For each simulation run the number of haplotypes being reconstructed is estimated via our model selection step as explained in Section S1-3. All the results discussed in this section are simulated with the parameters introduced in Section S2 (s = 0.05, 150 generations of E&R where allele frequencies are available every 10 generations, one locus carries the beneficial allele in three individual haplotypes, genotypes from the founder



S2a *Structure*

S2b *Frequency*

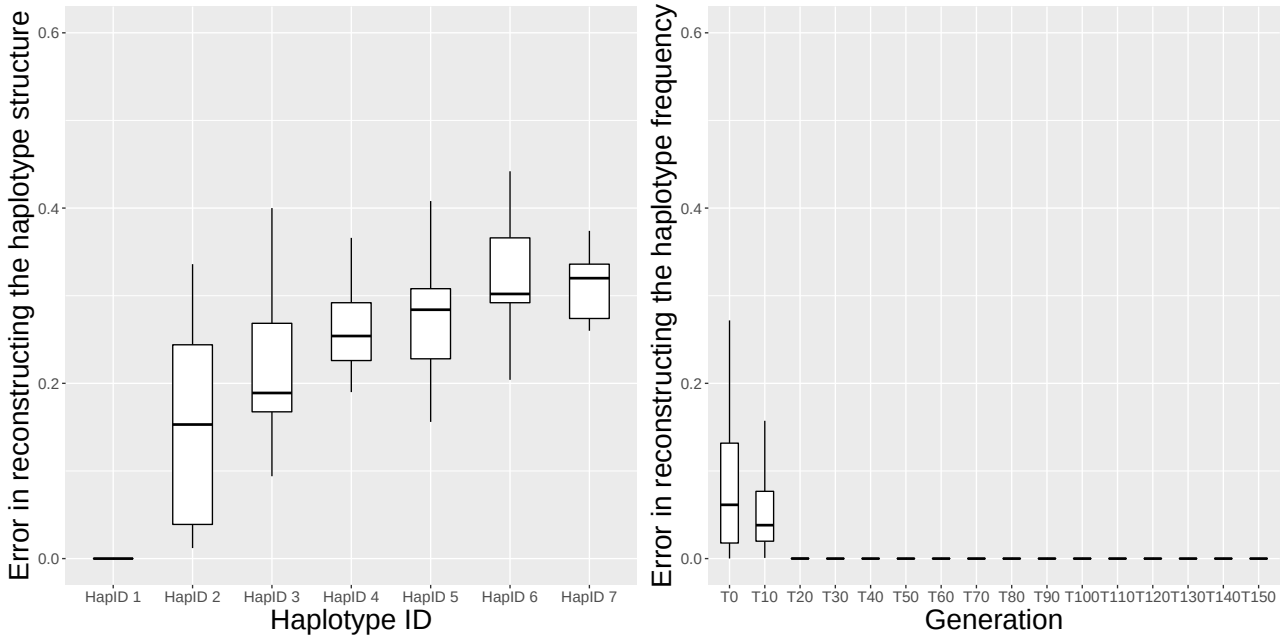
Figure S2: Haplotype reconstruction error for our basic selection scenario with *Drosophila simulans* based on 100 simulation runs. (a) Proportion of wrongly classified SNPs for each reconstructed haplotype. The haplotypes are displayed in decreasing order according to the frequency at the last time point. (b) Absolute difference between the true and estimated haplotype frequencies for each time point at which sequencing information is available.

population used in [Barghi et al., 2019]). Fig. S4 shows the accuracy depending on the selection pressure. As we expect, the error decreases when the selection pressure increases. We can observe that the effect is very pronounced for the experimental designs with large population size. This is because the reconstruction results become more and more accurate as the changes in haplotype frequency throughout time increase. When the populations size is small (e.g. in experiments using bigger organisms like mice), these haplotype frequency changes can occur under neutrality as well.

Our method requires information from multiple sources, which for E&R experiments correspond to sequenced time points. The number of time points at which the sequencing data are available mainly depends on the time and costs allocated to the experiment. As it is shown in the lower panel of Fig. S5 (and with a less pronounced effect in the upper panel), four time points do not contain enough information for any experimental design to obtain satisfactory results. However when the number of time points increases the error drops and this is consistent for all three experimental designs as well. It is also important to notice that the number of haplotypes we can reconstruct is smaller or equal to the number of available time points. This can also influence the power of our method under certain experimental designs where a high number of haplotypes is needed to capture the true dynamic of the haplotype frequencies in the given experiment.

In Fig. S6 we consider different numbers of haplotypes sharing the same beneficial allele. The more haplotypes share the same selective advantage, the less accurate the reconstruction becomes, unless the experiment is run for enough time to resolve the competition. If the competition is resolved and one or few haplotype(s) prevail, the reconstruction can reach high accuracy, however.

When looking at Fig. S7 we can see that a coverage of 5 is too low for accurate pooled allele frequency estimates. Thus our method cannot provide good estimates. When the coverage



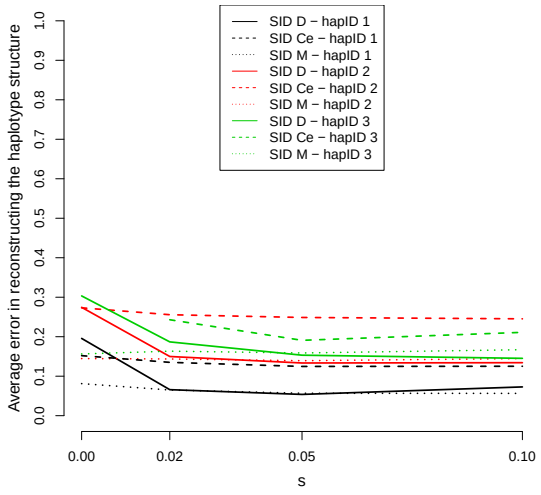
S3a *Structure*

S3b *Frequency*

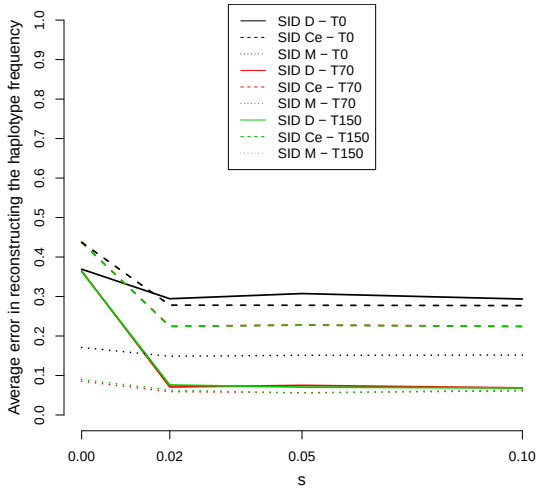
Figure S3: Haplotype reconstruction error for our basic selection scenario with *C. elegans* based on 100 simulation runs. (a) Proportion of wrongly classified SNPs for each reconstructed haplotype. The haplotypes are displayed in decreasing order according to the frequency at the last time point. (b) Absolute difference between the true and estimated haplotype frequencies for each time point at which sequencing information is available.

increases above $\lambda = 20$, not much accuracy is gained anymore. For our considered designs, more time points will be more beneficial than more reads in terms of accuracy. Compare for example, the results from our three experimental designs with fewer time points (e.g. 4) and high coverage ($\lambda = 80$) from Fig. S5 against those with more time point (16) and low coverage (e.g. $\lambda = 20$) from Fig. S7.

The last parameter we considered is the number of different haplotypes in the founder population (Fig. S8). Our simulations do not show a clear trend here. An intermediate number of haplotypes relative to the population size often seems to lead to the highest accuracy, this may be since in this case some - but not all- of the beneficial haplotypes tend to get lost by drift.

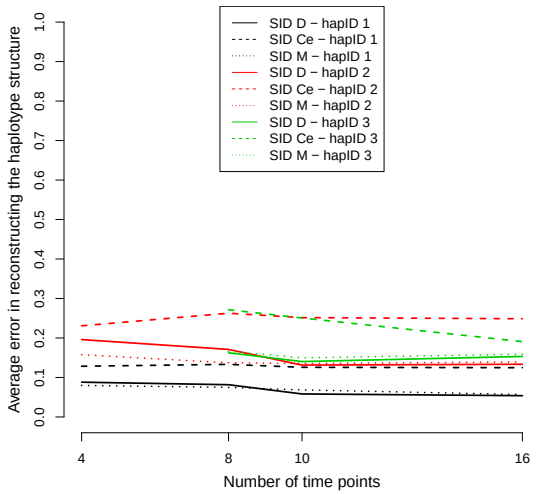


S4a *Structure*

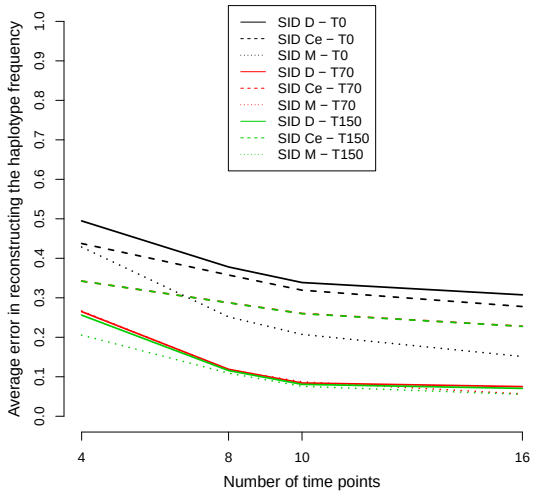


S4b *Frequency*

Figure S4: Dependence of the quality of our reconstruction approach on the selection coefficient. Simulation setup: $s \in 0, 0.02, 0.05, 0.1$ and all the other parameters as in Section S2. Results for *D. simulans* (solid lines), *C. elegans* (dashed lines), and the mice experiment (dotted lines) are shown. (a) Error in reconstructing the haplotype structure versus different values of the selection coefficient. For each experimental design, results for the three most frequent haplotypes are shown: hapID 1 (black lines), hapID 2 (red lines), and hapID 3 (green lines). (b) Error in reconstructing the haplotype frequencies versus different values of the selection coefficient. For each experimental design, results for time points T0 (black lines), T70 (red lines), and T150 (green lines) are shown.

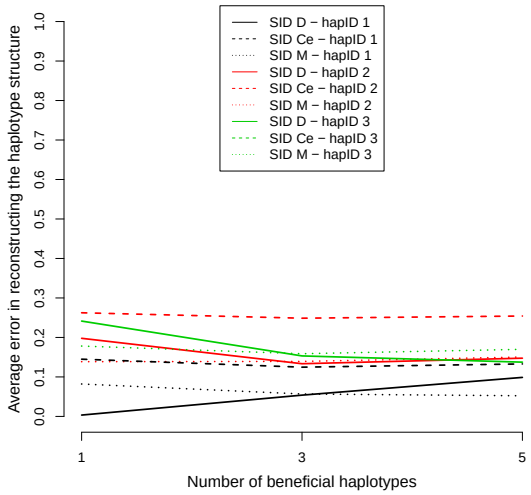


S5a *Structure*

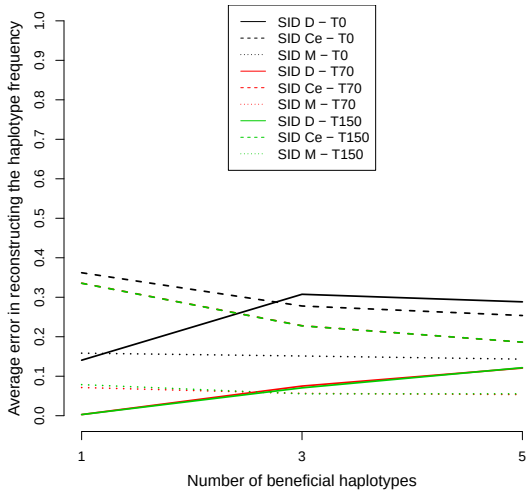


S5b *Frequency*

Figure S5: Dependence of the quality of our reconstruction approach on the number of sequenced time points. Results for *D. simulans* (solid lines), *C. elegans* (dashed lines), and the mice experiment (dotted lines) are shown. (a) Error in reconstructing the haplotype structure versus different numbers of sequenced time points. For each experimental design, results for the three most frequent haplotypes are shown: hapID 1 (black lines), hapID 2 (red lines), and hapID 3 (green lines). (b) Error in reconstructing the haplotype frequencies versus different numbers of sequenced time points. For each experimental design, results for time points T0 (black lines), T70 (red lines), and T150 (green lines) are shown.

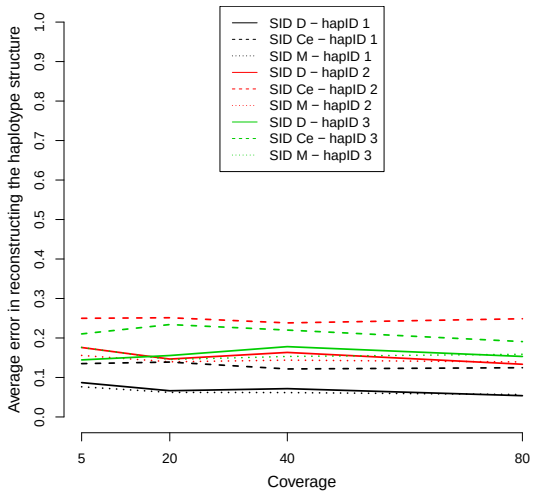


S6a *Structure*

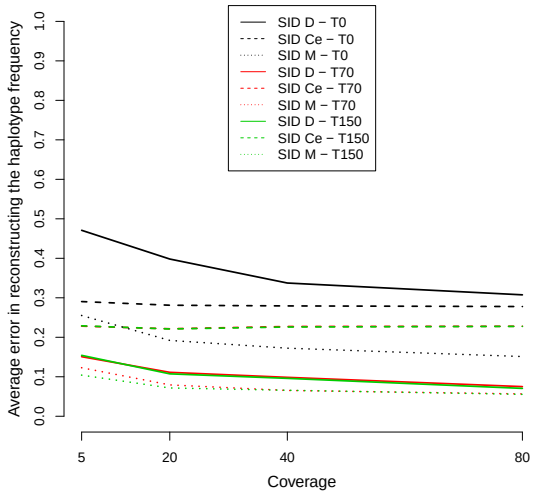


S6b *Frequency*

Figure S6: Dependence of the quality of our reconstruction approach on the number of haplotypes sharing the beneficial allele. Simulation setup: Number of haplotypes sharing the beneficial allele $\in 1, 3, 5$ and all the other parameters as in Section S2. Results for *D. simulans* (solid lines), *C. elegans* (dashed lines), and the mice experiment (dotted lines) are shown. (a) Error in reconstructing the haplotype structure versus different numbers of haplotypes sharing the beneficial allele. For each experimental design, results for the three most frequent haplotypes are shown: hapID 1 (black lines), hapID 2 (red lines), and hapID 3 (green lines). (b) Error in reconstructing the haplotype frequencies versus different numbers of haplotypes sharing the beneficial allele. For each experimental design, results for time points T0 (black lines), T70 (red lines), and T150 (green lines) are shown.

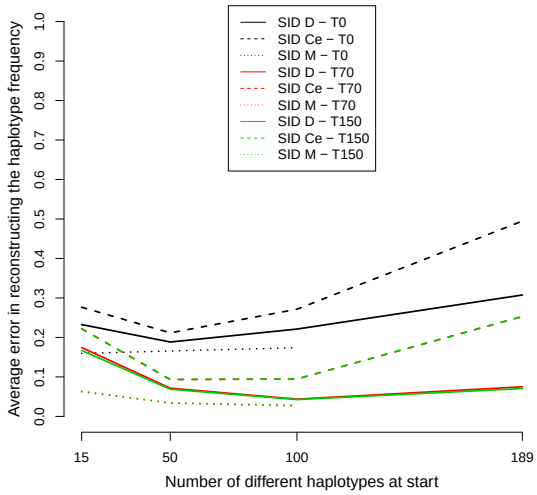


S7a *Structure*

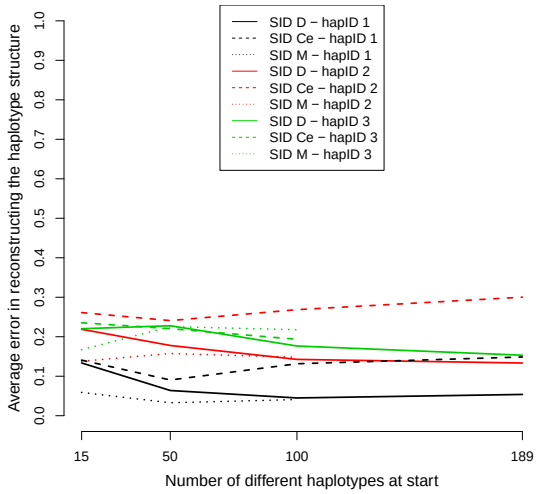


S7b *Frequency*

Figure S7: Dependence of the quality of our reconstruction approach on the mean coverage value λ. Simulation setup: $\lambda \in 5, 20, 40, 80$ and all the other parameters as in Section S2. Results for *D. simulans* (solid lines), *C. elegans* (dashed lines), and the mice experiment (dotted lines) are shown. (a) Error in reconstructing the haplotype structure versus different values of λ. For each experimental design, results for the three most frequent haplotypes are shown: hapID 1 (black lines), hapID 2 (red lines), and hapID 3 (green lines). (b) Error in reconstructing the haplotype frequencies versus different values of λ. For each experimental design, results for time points T0 (black lines), T70 (red lines), and T150 (green lines) are shown.



S8a *Structure*



S8b *Frequency*

Figure S8: Dependence of the quality of our reconstruction approach on the number of different haplotypes in the founder population. Simulation setup: Number of different haplotypes in the founder population $\in 15, 50, 100, 189$ and all the other parameters as in Section S2. Results for *D. simulans* (solid lines), *C. elegans* (dashed lines), and the mice experiment (dotted lines) are shown. (a) Error in reconstructing the haplotype structure versus different number of different haplotypes in the starting population. For each experimental design, results for the three most frequent haplotypes are shown: hapID 1 (black lines), hapID 2 (red lines), and hapID 3 (green lines). (b) Error in reconstructing the haplotype frequencies versus different number of different haplotypes in the starting population. For each experimental design, results for time points T0 (black lines), T70 (red lines), and T150 (green lines) are shown.

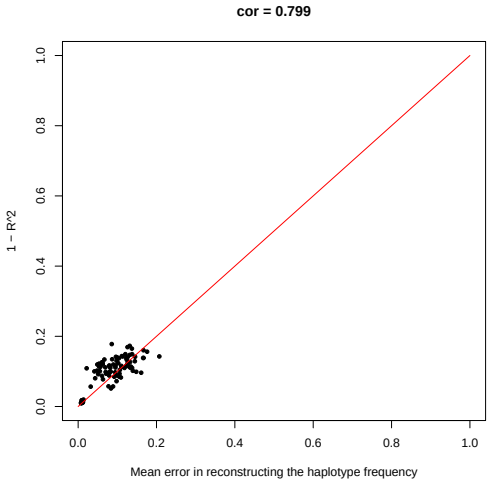
S4 Accuracy measures

When applying our method to real data the true haplotypes are unknown and the error cannot be assessed. For this reason, we provide measures of accuracy for the full reconstruction (namely R^2) for the haplotype structures and for the haplotype frequencies (see Section S1-4 for a more detailed explanation on how the scores are computed). To see how well these accuracy measures coincide with the actual amount of error, we provide simulation results for our three simple selection scenarios. We expect high scores when the error is low and vice-versa.

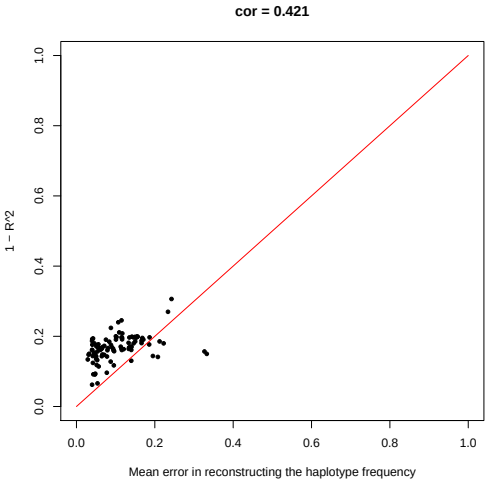
We plot R^2 against the overall error in the reconstruction of the haplotype frequency for our three simple selection scenarios in Fig. S9. This figure shows that for the scenario with small population size the correlation between R^2 and error is relatively high (0.799), however for large population sizes either the correlation is low (0.421) or the R^2 is underestimating our error in reconstruction (see Fig. S9c). When the correlation is low, the error is only slightly over estimated by R^2, whereas in the case of Fig. S9c we have a group of scenarios where the R^2 is too liberal. However, if we discard the scenarios where the haplotype frequency change of the most frequent reconstructed haplotype is small (< 0.1) then the correlation in Fig. S9b increases up to 0.521 and the scenarios where R^2 is underestimating the error in S9c are not included in the analysis anymore. If the frequency change of the dominant haplotype is small, it means that selection is either not present (neutral dynamic in a large population), or its signal cannot be captured by our method. Therefore we recommend to look at the combination of both R^2 and frequency change. This was the motivation for our filtering criteria proposed in Section S1-4.

Our structure specific stability score (see equation S7 in section S1-4) is also correlated with the error in the reconstructed haplotype configuration (see Figs. S10a, S11a, and S12a). The high correlation shows that this measure is useful in applications. To test our accuracy measure for the haplotype frequencies, we checked how often each true frequency is contained inside the accuracy interval. The results in Fig. S13 show a high match between our bands and the true haplotypes, especially for late time points. Histograms of band sizes for these three scenarios can be found in figures S10b, S11b, and S12b, and they reveal that the bands are usually quite small (about 50% or more of the observed bandwidth being smaller than 0.05 in the worst scenario). These results demonstrate that these scores are concordant with the actual errors.

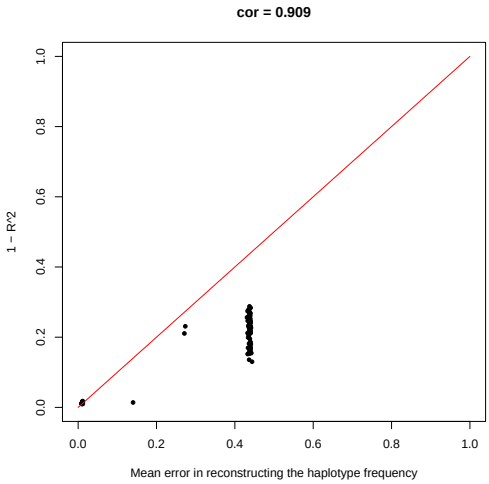
We recommend to use the haplotype specific stability intervals and stability scores after ensuring that our overall quality measures (R^2 and frequency change of the dominant haplotype) are good enough.



S9a *Longshank mice exp.*

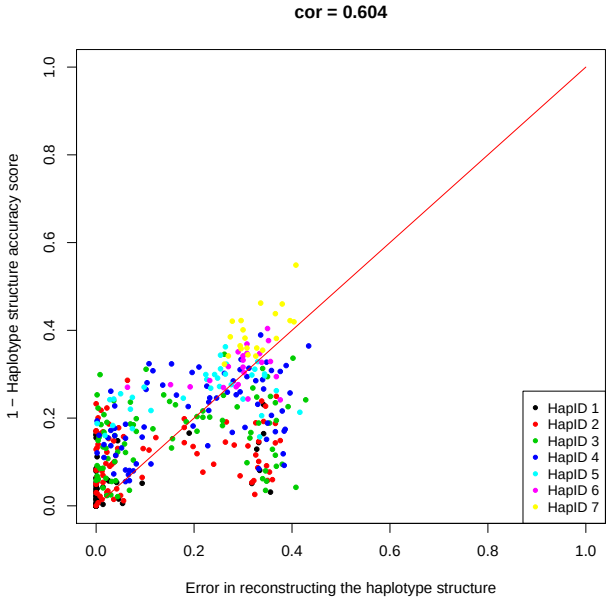


S9b *D. simulans*

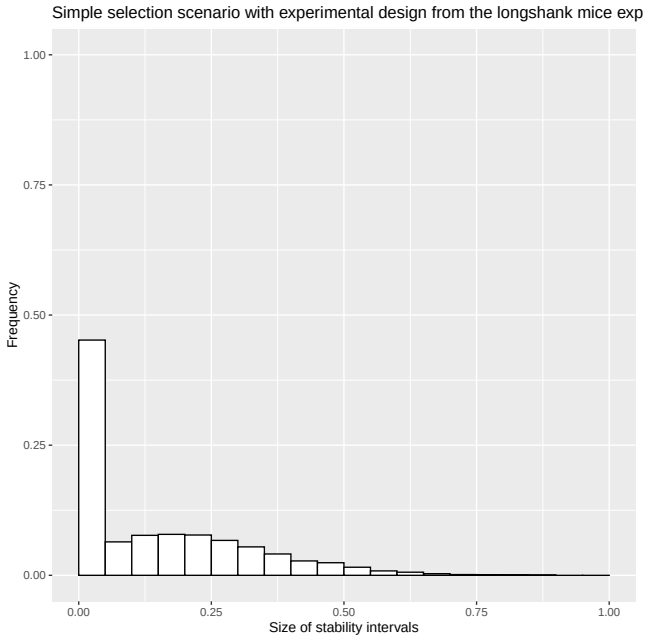


S9c *C. elegans*

Figure S9: Mean error in reconstructing the haplotype frequency versus $1 - R^2$ for (a) the Longshank mice experimental design, (b) the *D. simulans* experimental design, and (c) the *C. elegans* experimental design

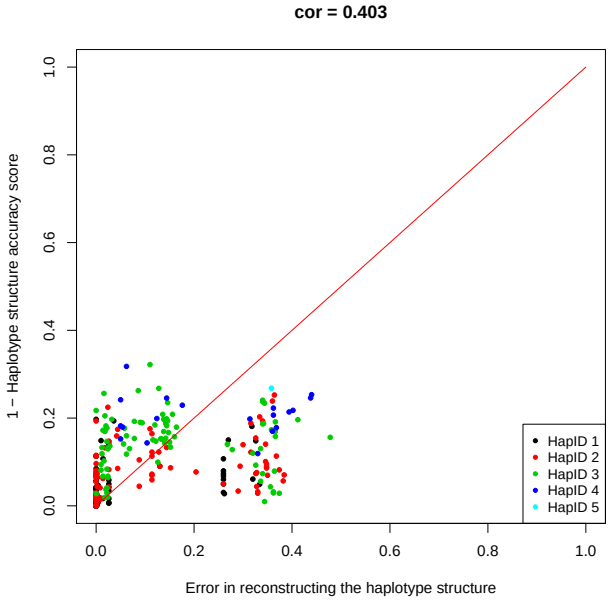


S10a *Structure*

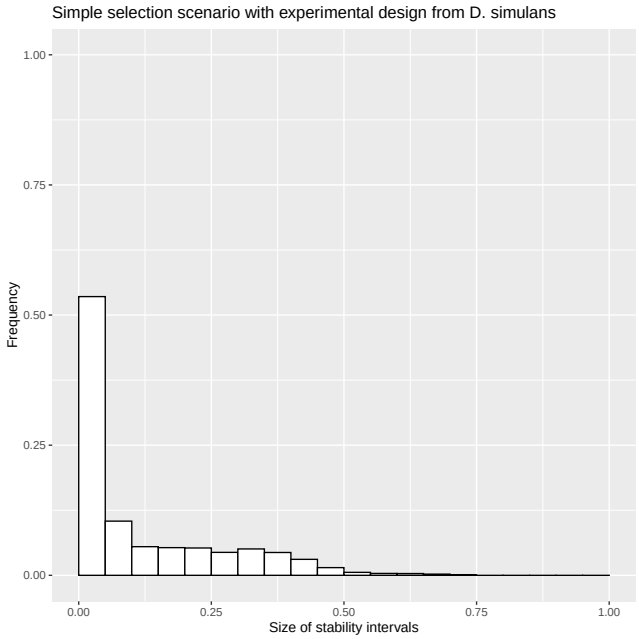


S10b *Frequency*

Figure S10: (a) Proportion of incorrectly estimated alleles when reconstructing the haplotype structure versus the corresponding accuracy scores for the Longshank mice experimental design. (b) Size of the accuracy intervals for the reconstructed haplotype frequency for the Longshank mice experimental design.

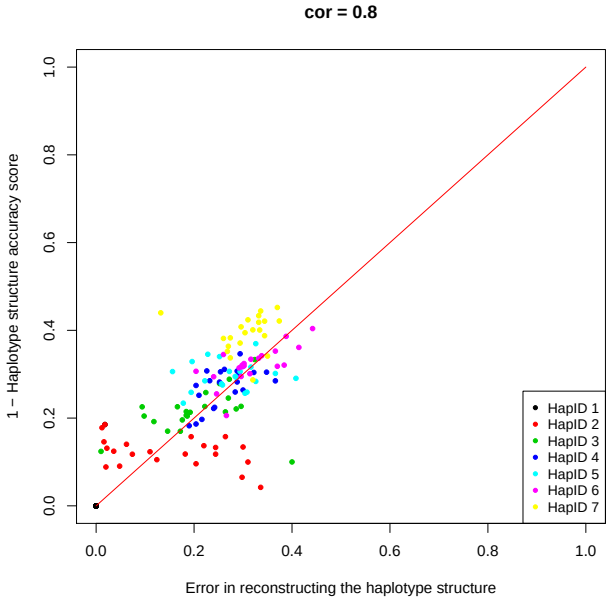


S11a *Structure*

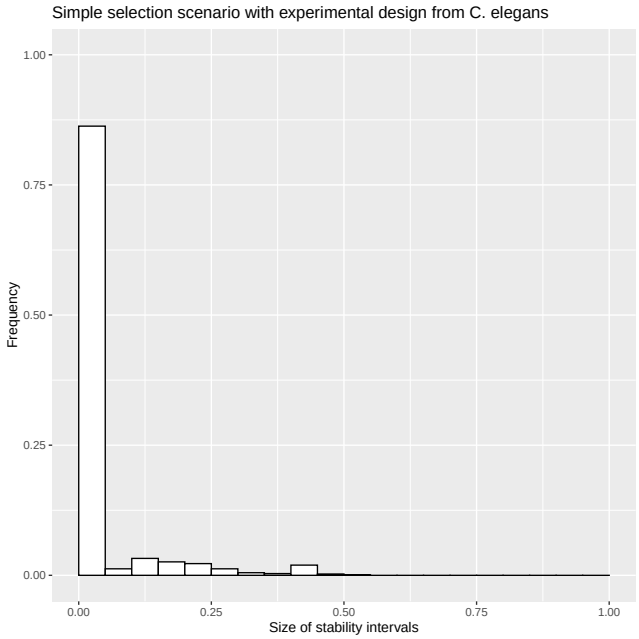


S11b *Frequency*

Figure S11: (a) Proportion of incorrectly estimated alleles when reconstructing the haplotype structure versus the corresponding accuracy scores for the *D. simulans* experimental design. (b) Size of the accuracy intervals for the reconstructed haplotype frequency for the *D. simulans* experimental design.



S12a *Structure*



S12b *Frequency*

Figure S12: (a) Proportion of incorrectly estimated alleles when reconstructing the haplotype structure versus the corresponding accuracy scores for the *C. elegans* experimental design. (b) Size of the accuracy intervals for the reconstructed haplotype frequency for the *C. elegans* experimental design.

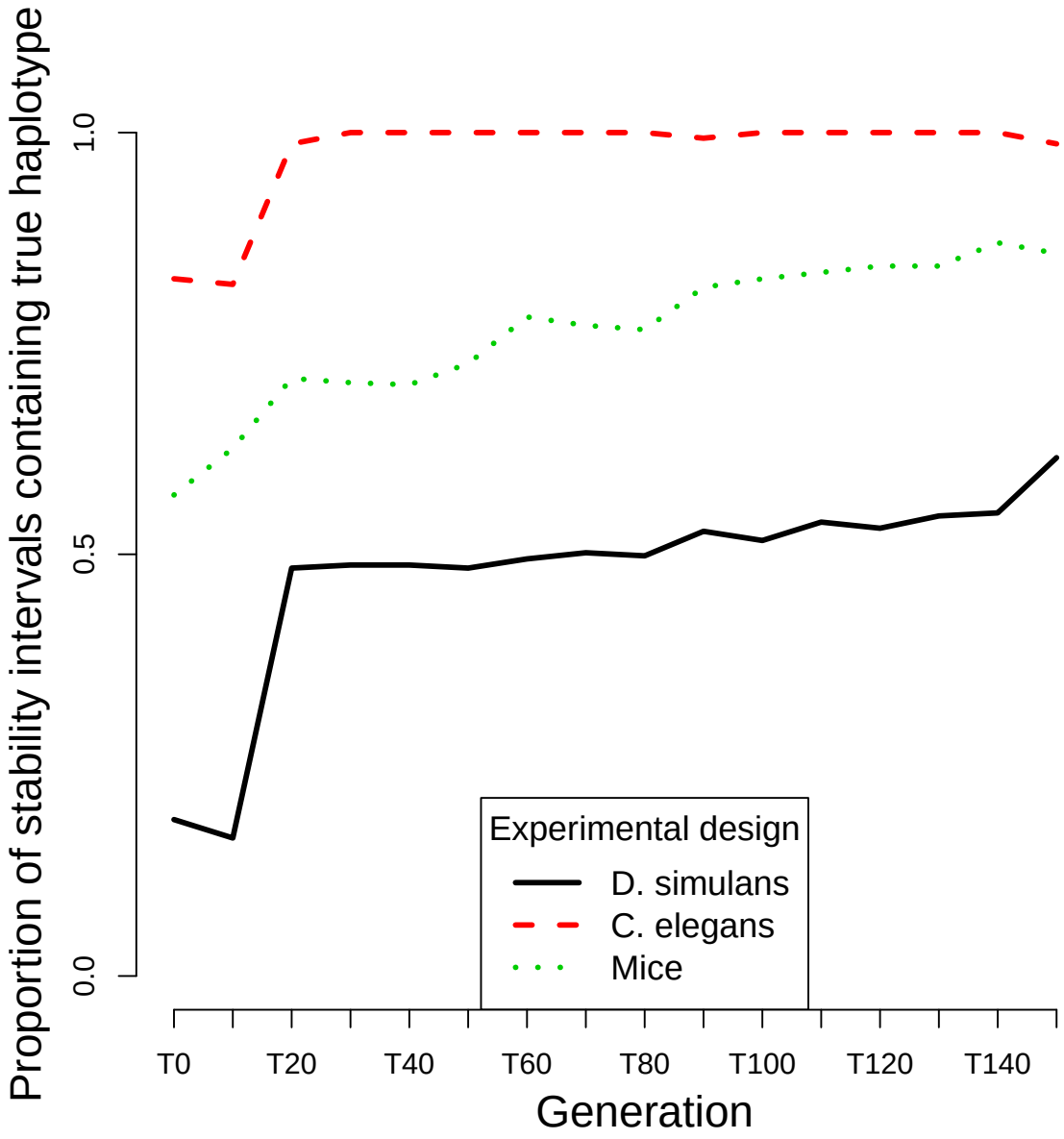
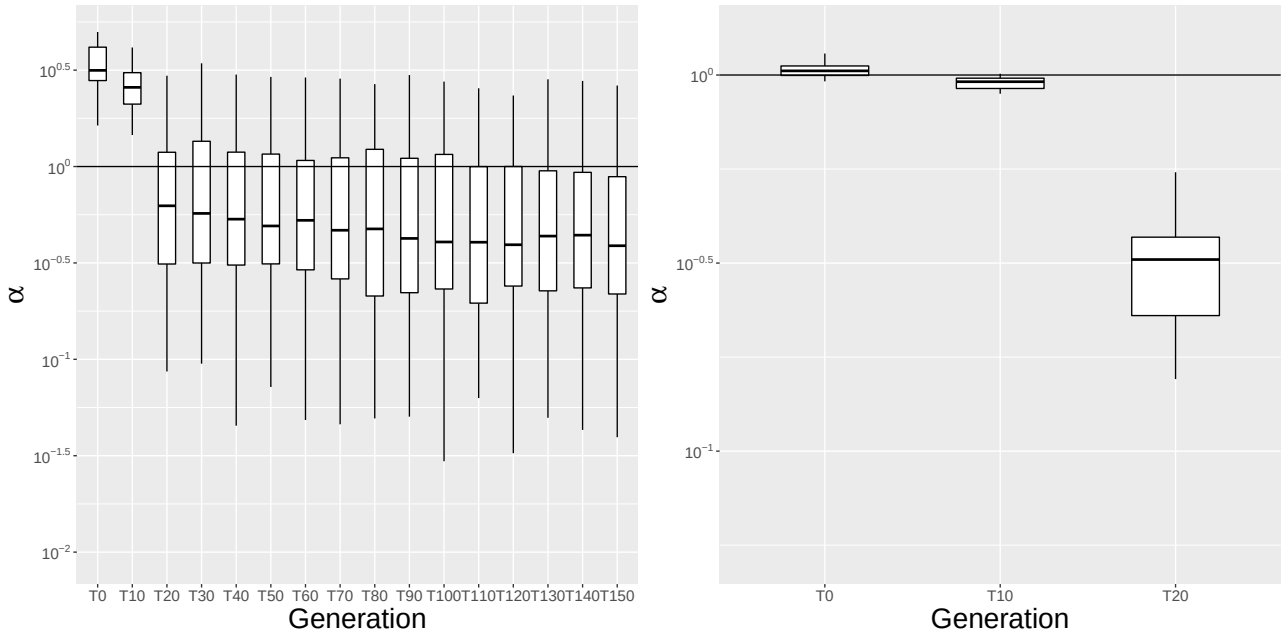


Figure S13: Proportion of stability intervals containing the true haplotype for our three simple selection scenarios.

S5 Improved allele frequency estimates: additional results



S14a *D. simulans*

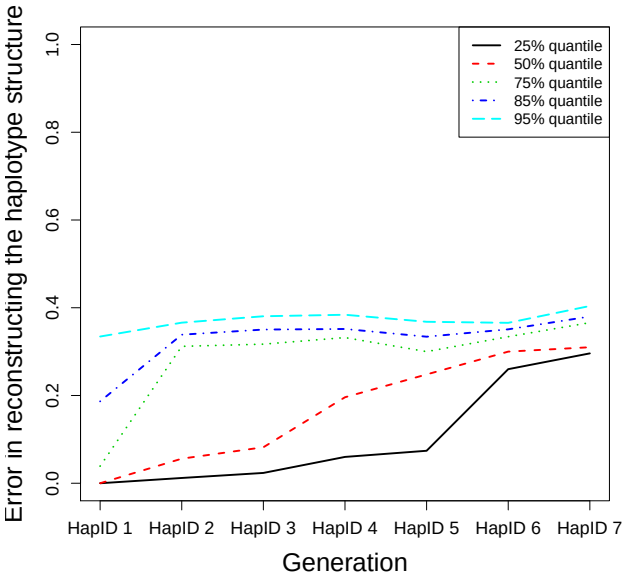
S14b *C. elegans*

Figure S14: Error ratio (α) between haplotype based allele frequency estimates (numerator) and the pool sequencing estimates (denominator) plotted on a log-scale. Results from 100 simulation runs based on the experimental designs in [Barghi et al., 2019] and [Noble et al., 2019].

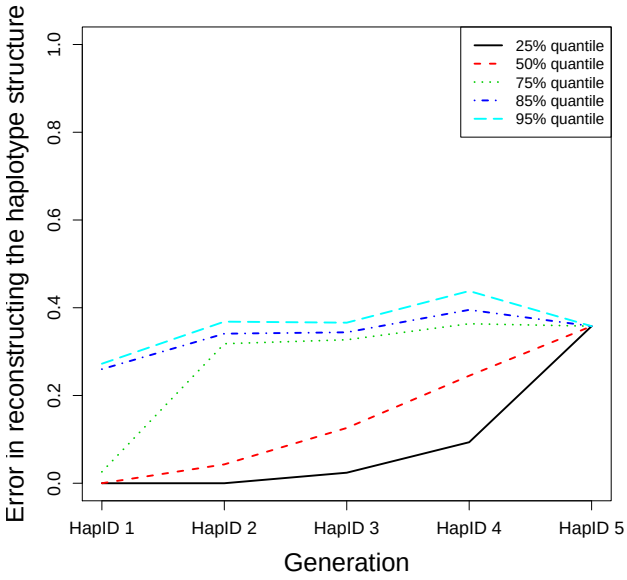
Late time points for the *C. elegans* example are not shown as both errors in reconstructing the allele frequency data are negligible and thus the ratio cannot be computed. Further information on the later time point can be found in Fig. S17c where all scenarios are included.

S6 Analysis of outliers

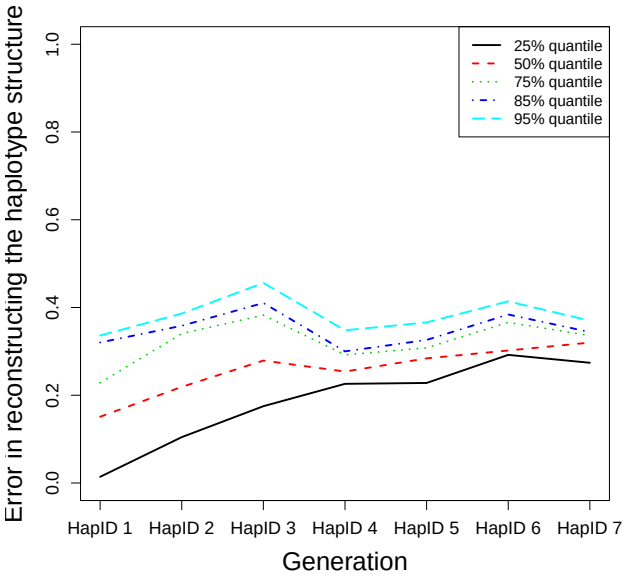
Here we consider all the simulation results for the three simple selection scenarios without filtering using R^2 and the frequency change of the haplotype with highest frequency. Figs. S15, S16, and S17 show the quantiles of the errors in reconstructing the haplotype frequency and structure and for α. The proportion of scenarios leading to outliers in the error measurements is $15\%, 19\%$, and 78% for the simulations based on the *Drosophila simulans*, Longshank mice, and *C. elegans* experimental design respectively. For *C. elegans* the proportion of outlier simulation runs is considerably higher than for the other two scenarios. Indeed, the population size in the *C. elegans* experiment is much larger than for the other organisms. When the dynamic is neutral in such a large population, there is a large number of haplotypes at very low frequency. These haplotypes are often aggregated within a few estimates at intermediate (and constant) frequency.



S15a *Longshank mice exp.*

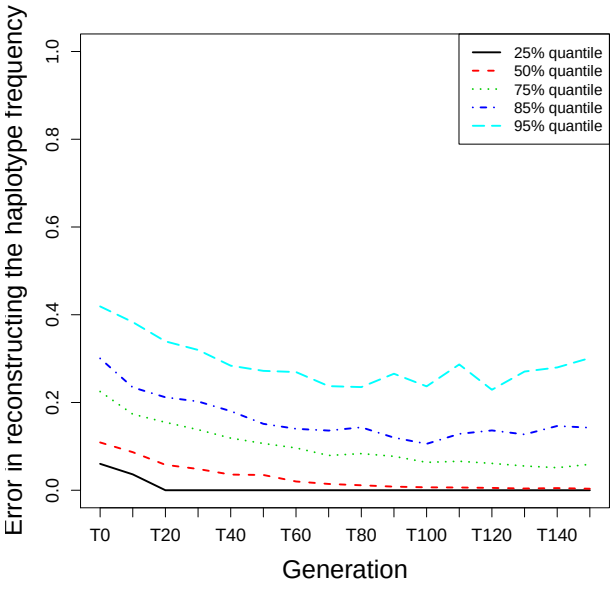


S15b *D. simulans*

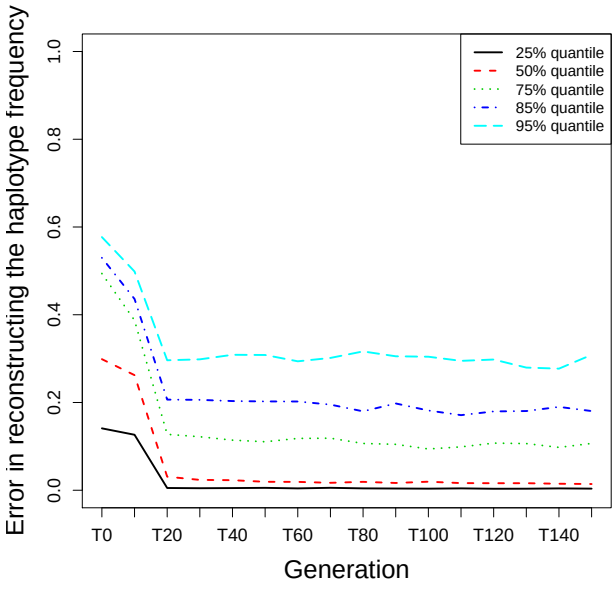


S15c *C. elegans*

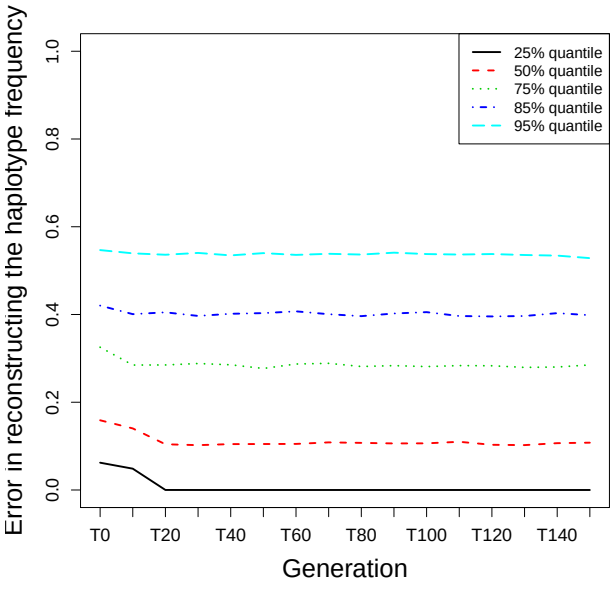
Figure S15: Quantiles of the error in reconstructing the haplotype structure for (a) the Longshank mice experimental design, (b) the *D. simulans* experimental design, and (c) the *C. elegans* experimental design.



S16a *Longshank mice exp.*

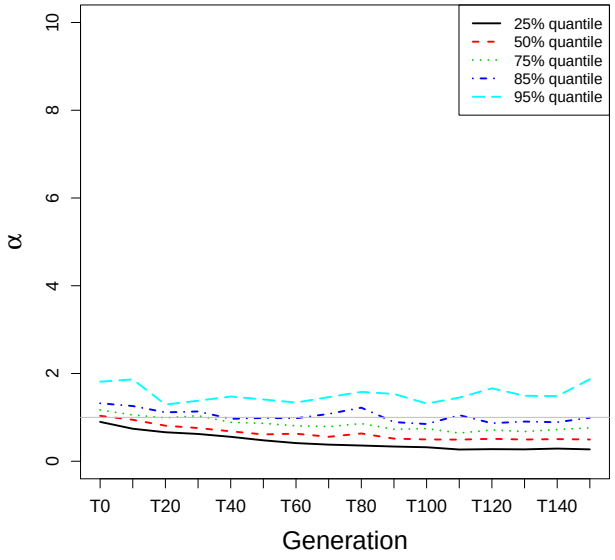


S16b *D. simulans*

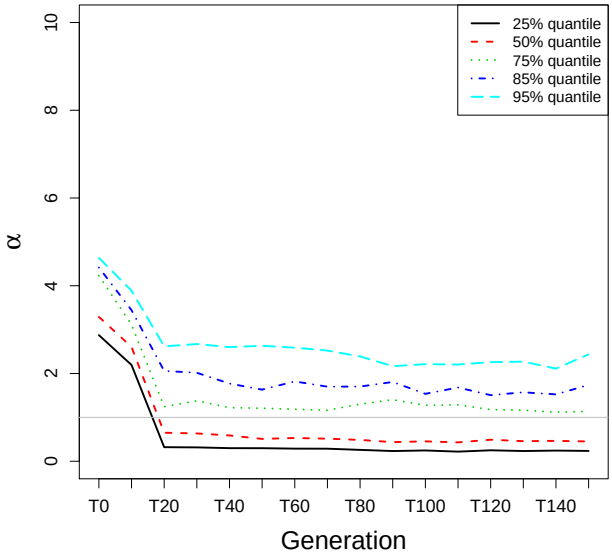


S16c *C. elegans*

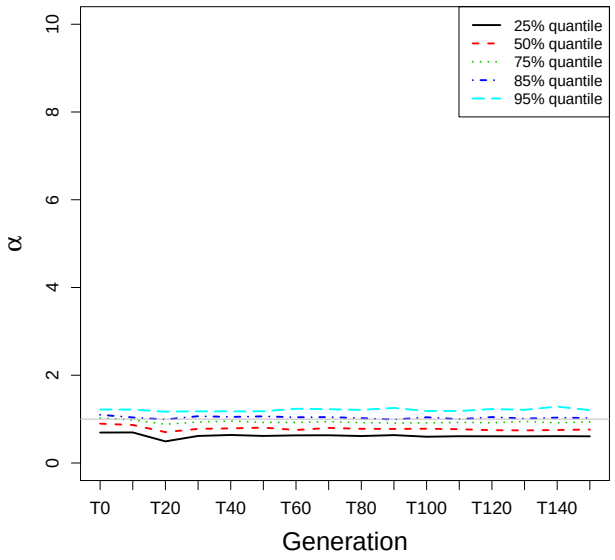
Figure S16: Quantiles of the error in reconstructing the haplotype frequency for (a) the Longshank mice experimental design, (b) the *D. simulans* experimental design, and (c) the *C. elegans* experimental design.



S17a *Longshank mice exp.*



S17b *D. simulans*



S17c *C. elegans*

Figure S17: Quantiles of the ratio between the error in estimating the allele frequencies from the reconstructing haplotypes versus pool sequencing (α) for (a) the Longshank mice experimental design, (b) the *D. simulans* experimental design, and (c) the *C. elegans* experimental design.

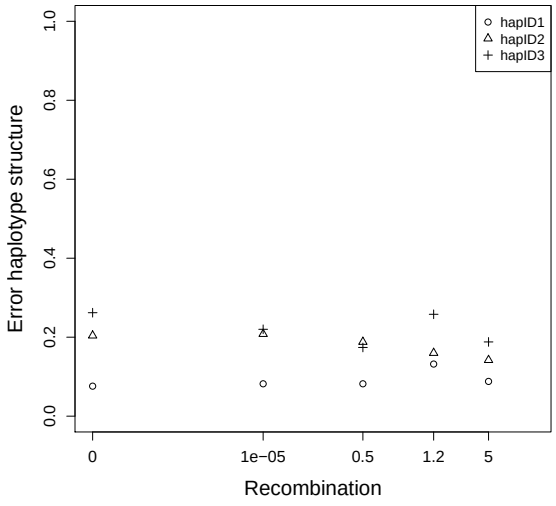
S7 Recombination

Here, we describe an example involving recombination simulated with MimicrEE2 [Vlachos and Kofler, 2018], following the *Drosophila simulans* setup. We investigate different values for the recombination rate, which is assumed homogeneous throughout the whole region. This example, is supposed to illustrate how recombination can affect our haplotype reconstruction. Results based on one simulation run are shown in Fig. S18. Recombination rate is in cM/Mb which is converted by MimicrEE2 to a lambda-value of a Poisson distribution using Haldane's map function.

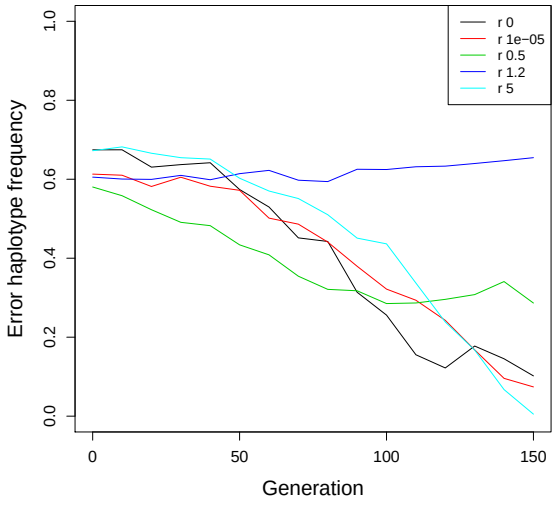
Overall, we see that recombination does not much affect the quality of the reconstructed haplotype structure. On the other hand, the effect on the estimated frequencies can be complex. In principle, we expect that the reconstruction becomes more difficult with high recombination rates, as new haplotypes will constantly arise. Thus the number of haplotypes having low frequency will not decrease during the experiment. As they cannot be reconstructed, their frequencies will be partially attributed to other haplotypes. Indeed, if recombination events occur close to the boundaries of the considered DNA segment, the new haplotypes will be almost identical to the two original haplotypes.

However, in cases such as the simulation with r = 5, we observe that recombination can lead to a new more beneficial haplotype rising considerably in frequency which can then be reconstructed with high accuracy. This can be seen in Figs. S18 (right panel) and S19.

An obvious approach to avoid potential problems arising from recombination is to choose the window size small enough, so that only few recombination events will occur during the experiment.



S18a *Structure*



S18b *Frequency*

Figure S18: (a) Proportion of wrongly classified SNPs for each reconstructed haplotype for different values of the recombination rate. The haplotypes are displayed in decreasing order according to the frequency at the last time point (b) Absolute difference between the true and estimated haplotype frequencies for each time point at which sequencing information is available. This plot includes only the first dominant haplotype. Different colours indicate different recombination rates.

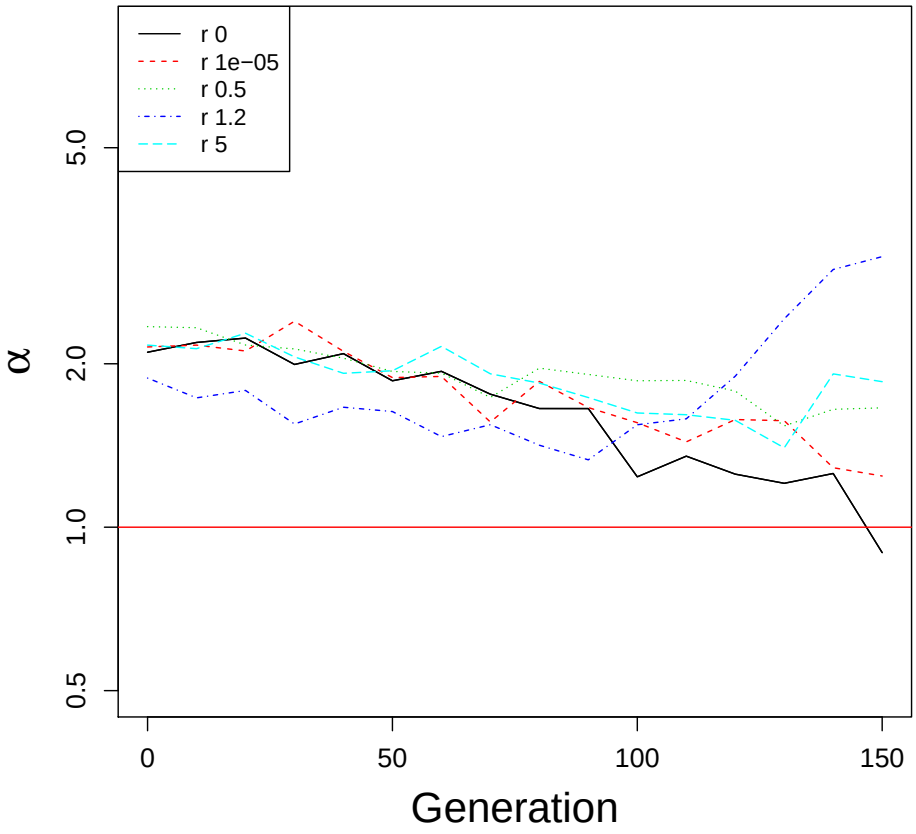


Figure S19: Error ratio (α) between haplotype based allele frequency estimates (numerator) and the pool sequencing estimates (denominator) plotted on a log-scale for different values of the recombination rate.

S8 Validation of our results using read data

We used read data from [Barghi et al., 2019] as a further validation of our reconstructed haplotypes. These data are provided by the authors after the reads were trimmed and mapped to the genome and after duplicates have been removed. These steps, as well as the DNA extraction and library preparation are described in [Barghi et al., 2019]. In order to be consistent with the allele frequency data and thus with the reconstructed haplotypes we only used SNPs analysed in the original paper. Furthermore, as in [Barghi et al., 2019] for a given SNP we kept the information from the reads only when the respective base quality score was higher than 20. As in Section 5, for this analysis we chose a region under selection according to the p-values from the modified chi-squared test in [Spitzer et al., 2020]. Here, we considered the region from 11.239636 to 11.733131 Mb of chromosome 2L in replicate three. All comparisons with the reads are performed at generation 60.

For each read partially overlapping the region of interest we apply the following steps. First, we combined paired end reads to a long sequence with a missing part in the middle because read pairs belong to the same haplotype. Then, we polarize the set of read data for the rising allele, as we did for the allele frequency data.

In order to compare the read data with the reconstructed haplotypes, we considered sliding windows of 1000 SNPs and performed the following analysis on each window. For our first comparison, we selected the most similar read for each reconstructed haplotype and window. Fig. S20 shows the proportion of mismatches between haplotype and corresponding read without considering missing data. From the example we can see that most haplotypes have a good match with the reads, which is a further validation of the fact that the haplotype structure we reconstruct with our method is accurate. However, the number of positions entering this comparison for each read is limited (between 32 and 59). Indeed, there are always many missing values in each read as read length is limited and they might not overlap a region entirely and genomic positions might be filtered out for low base quality scores.

We decided then to examine these results in terms of haplotype frequency as well. Because reads are short and insert sizes generate missing values, we cannot compare the frequencies of the reads with those of the haplotypes directly. At the same time, using single SNPs would not be informative in this situation because we already validated the power of our method in reconstructing allele frequency data (see Section 4.2). Thus, we decided to consider the smallest available linked unit, and we performed our comparison on pairs of subsequent SNPs using the frequencies of the four possible genotypes of each pair.
The results from this comparison are shown in Fig. S21. From these examples we can see that also the frequency of the pairs of SNPs are estimated with low error from our reconstructed haplotypes, which strongly suggests that the reconstructed haplotypes capture the signal from the true haplotypes in the population correctly.

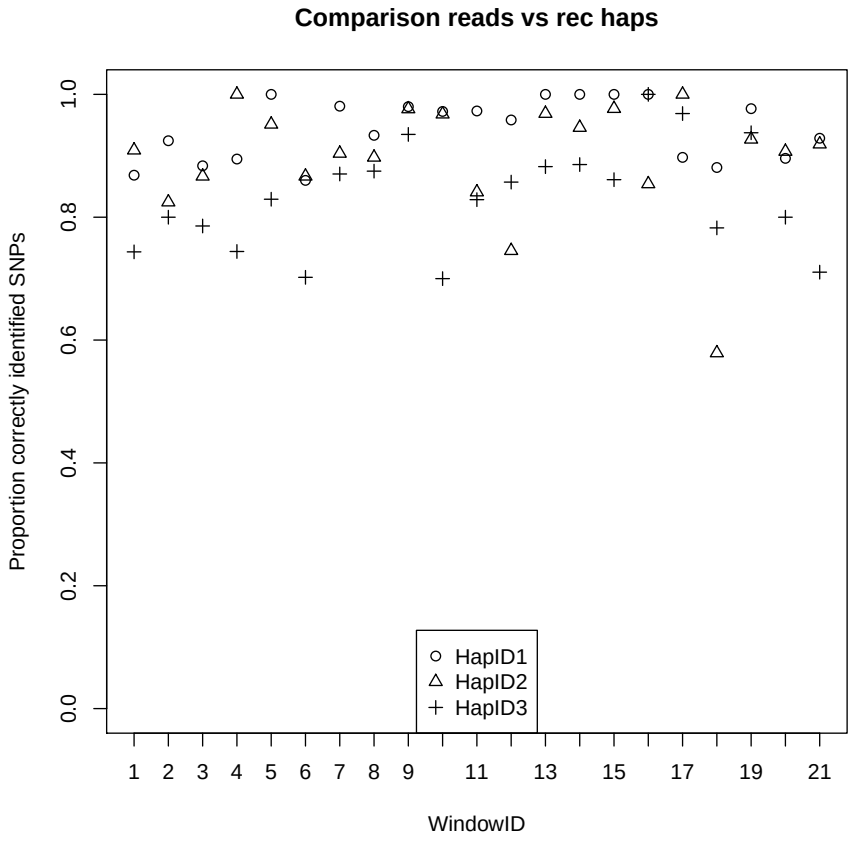


Figure S20: Comparison of the reconstructed haplotype structure with the read data.

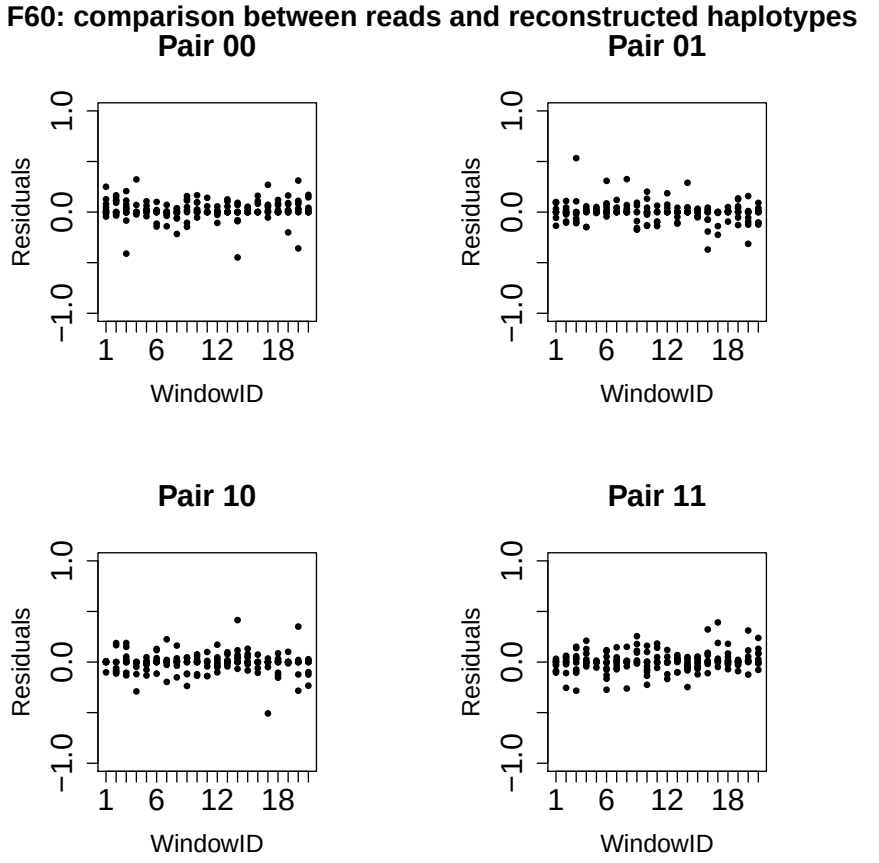
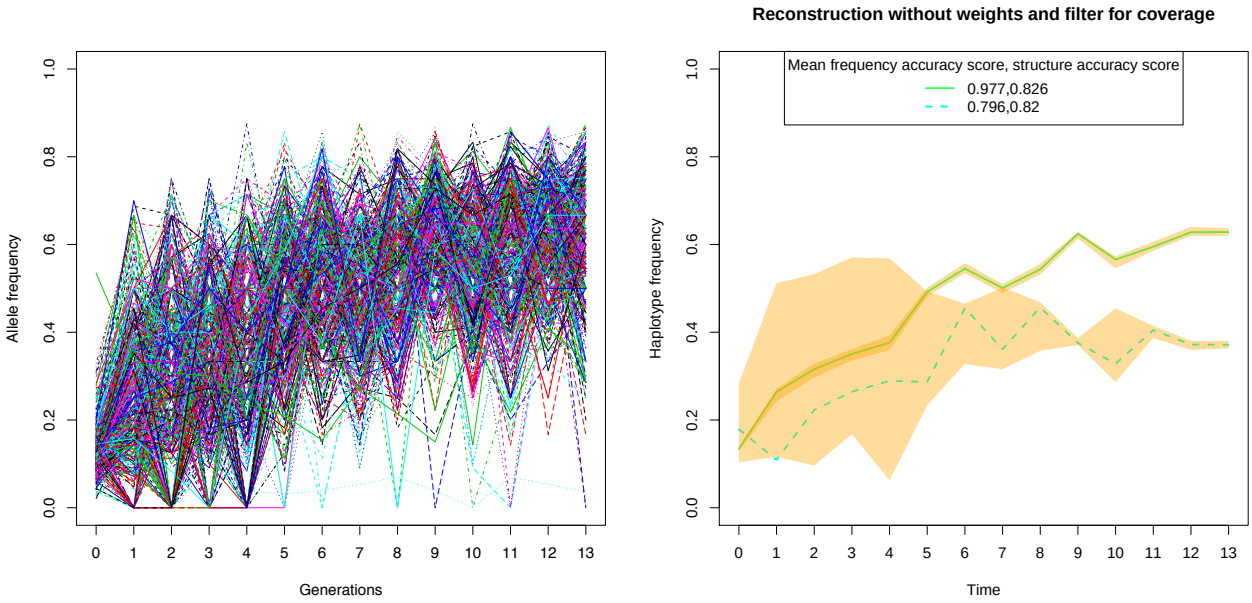


Figure S21: Residuals of the estimated frequency of pairs of SNPs from read data versus the estimated frequency of pairs of SNPs from reconstructed haplotypes.

S9 Results for the Longshank mice experiment

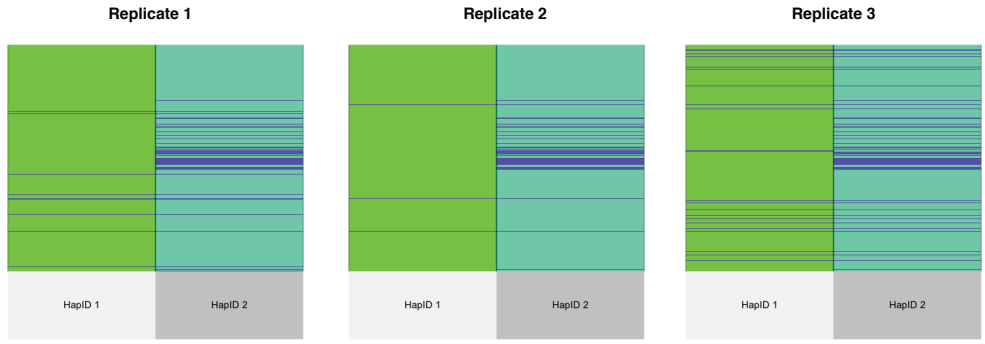


S22a *Alleles*

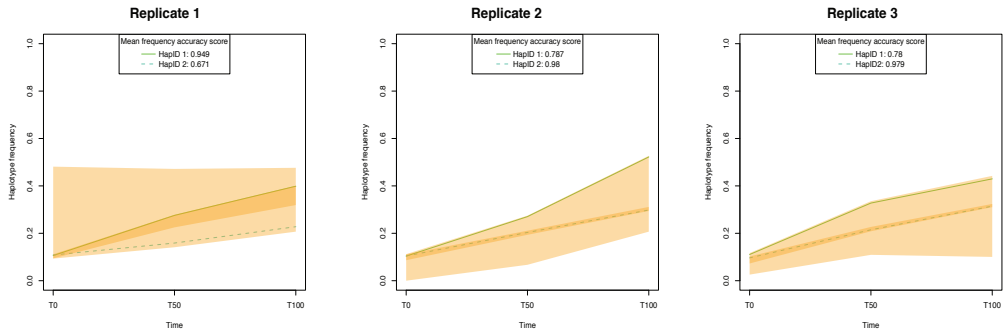
S22b *Haplotypes*

Figure S22: (a) Observe time-series of allele frequencies. (b) Reconstructed haplotype frequencies with accuracy intervals (in yellow) and mean accuracy scores.

S10 Additional results from the *C. elegans* data set from [Noble et al., 2019]

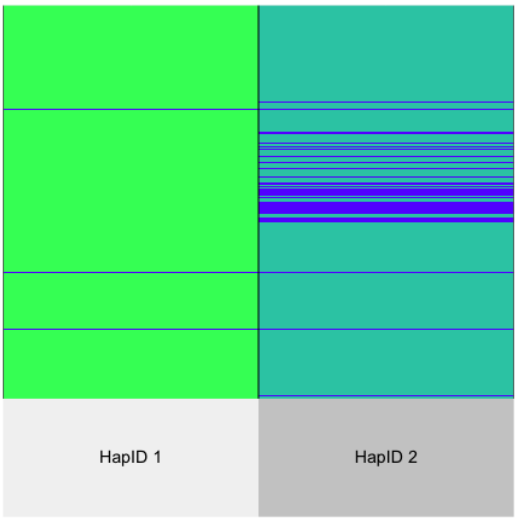


S23a *Structure*

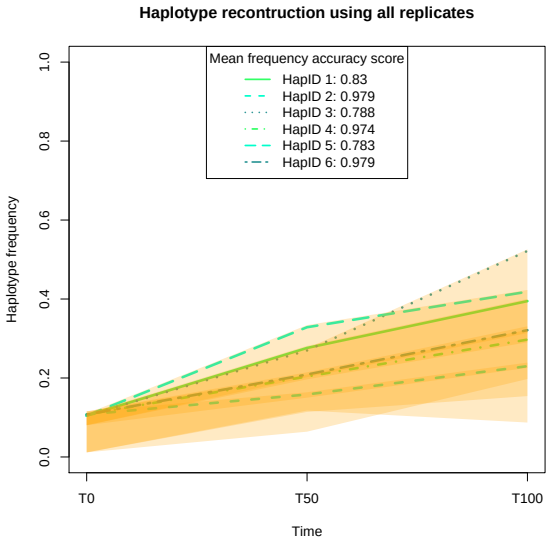


S23b *Frequency*

Figure S23: Haplotype reconstruction for data from [Noble et al., 2019] (a) Match between the haplotype structure reconstructed from the allele frequency data and the sequenced founder haplotypes. Blue lines indicate mismatch positions. (b) Reconstructed haplotype frequencies with accuracy intervals (in yellow) and mean accuracy scores.



S24a *Structure*



S24b *Frequency*

Figure S24: (a) Match between reconstructed haplotype structure and sequenced founder haplotypes using all the three replicates from [Noble et al., 2019] at the same time. Blue lines indicate mismatches. (b) Reconstructed haplotype frequencies with accuracy intervals (in yellow) and mean accuracy scores using all the three replicates from [Noble et al., 2019] at the same time.

References

[Barghi et al., 2019] 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):e3000128.

[Behr et al., 2018] Behr, M., Holmes, C., and Munk, A. (2018). Multiscale blind source separation. *The Annals of Statistics*, 46(2):711–744.

[Behr and Munk, 2017] Behr, M. and Munk, A. (2017). Identifiability for Blind Source Separation of Multiple Finite Alphabet Linear Mixtures. *IEEE Trans. Information Theory*, 63(9):5506–5517.

[Castro et al., 2019] 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.

[Diamantaras and Chassioti, 2000] Diamantaras, K. I. and Chassioti, E. (2000). Blind separation of n binary sources from one observation: A deterministic approach. In *International Workshop on Independent Component Analysis and Blind Signal Separation*, pages 93–98, Helsinki.

[Efron, 1979] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

[Gavish and Donoho, 2014] Gavish, M. and Donoho, D. L. (2014). The Optimal Hard Threshold for Singular Values is 4/\sqrt 3. *IEEE Transactions on Information Theory*, 60(8):5040–5053.

[Griffin et al., 2017] Griffin, P. C., Hangartner, S. B., Fournier-Level, A., and Hoffmann, A. A. (2017). Genomic trajectories to desiccation resistance: Convergence and divergence among replicate selected *Drosophila* lines. *Genetics*, 205(2):871–890.

[Jónás et al., 2016] Jónás, A., 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.

[Kraaijeveld and Godfray, 2008] Kraaijeveld, A. R. and Godfray, H. C. (2008). Selection for resistance to a fungal pathogen in Drosophila melanogaster. *Heredity*, 100(4):400–406.

[Noble et al., 2019] Noble, L. M., Rockman, M. V., and Teotónio, H. (2019). Gene-level quantitative trait mapping in an expanded multiparent experimental evolution panel. *bioRxiv preprint 589432; doi: https://doi.org/10.1101/589432*.

[Schlötterer et al., 2014] 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.

[Spitzer et al., 2020] Spitzer, K., Pelizzola, M., and Futschik, A. (2020). Modifying the Chi-square and the CMH test for population genetic inference: Adapting to overdispersion. *The Annals of Applied Statistics*, 14(1):202–220.

[Vlachos and Kofler, 2018] Vlachos, C. and Kofler, R. (2018). MimicrEE2: Genome-wide forward simulations of Evolve and Resequencing studies. *PLoS Computational Biology*, 14(8).

[Waples, 1989] Waples, R. S. (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics*, 121:379–391.

[Weber, 1996] Weber, K. E. (1996). Large genetic change at small fitness cost in large populations of Drosophila melanogaster selected for wind tunnel flight: Rethinking fitness surfaces. *Genetics*, 144(1):205–213.