Multiple Haplotype Reconstruction from Allele Frequency Data

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

1 Abstract

We propose a new method that is able to accurately infer major haplotypes and their frequencies just from multiple samples of allele frequency data. Our approach seems to be the first that is able to estimate more than one haplotype given such data. Even the accuracy of experimentally obtained allele frequencies can be improved by re-estimating them from our reconstructed haplotypes.
Reconstructing haplotypes from sequencing data is of interest to several areas of biological and medical research. In evolutionary genetics, for instance, haplotypes help to better understand the genetic architecture of adaptation. Here we consider genomic time series data from three evolve and re-sequence experiments as an application. However, the approach can in principle be used in a wider context, as only data from multiple samples are needed, not necessarily collected over time.

2 Introduction

Understanding the haplotype composition of populations can frequently provide crucial information in studies relying on genetic data. Such investigations can be on different topics, such as the identification of genetic associations with diseases [Tewhey et al., 2011], the imputation of missing genotype data [Marchini et al., 2007], the inference of demographic population histories [Tishkoff et al., 1996], and the detection of traces of selection [Sabeti et al., 2002]. Therefore several recently proposed methods aim at extracting haplotypes (e.g. [Delaneau et al., 2019], [Browning et al., 2018], [Loh et al., 2016]) from sequencing data. A review of further methods and applications can be found in [Browning and Browning, 2011].
For human populations, great efforts have been put into sequencing large numbers of individuals accompanied by further efforts in the development of fast algorithms to obtain haplotype information by phasing the read data.
In other fields of applications, however, resources for sequencing are more scarce. In studies on the genetic basis of adaptation of non-human populations, for instance, the available haplotype information is often very limited, or lacking entirely. To lower the cost of the experiment, populations are frequently sequenced as a pool [Burke et al., 2010], [Illingworth et al., 2012],

[Barghi et al., 2019]. This approach provides genome-wide allele frequency data on a SNP level [Futschik and Schlötterer, 2010], [Schlötterer et al., 2014], but does not lead to any direct haplotype information. It is frequently used in Evolve and Resequence (E&R) experiments [Turner et al., 2011]. These experiments implement experimental evolution, with one or more (replicate) populations being followed for several generations in the lab under artificial selection and sequenced multiple times to obtain time series of allele frequency data.

Efforts have been made to infer haplotypes and their frequencies given allele frequency data. The methods by [Excoffier and Slatkin, 1995], [Pirinen, 2009], [Gasbarra et al., 2011], [Long et al., 2011], [Kessner et al., 2013] and [Cao and Sun, 2015] use known founder haplotypes to estimate the frequency trajectories of these haplotypes over time. The approach proposed in [Franssen et al., 2017], and optimized in [Otte and Schlötterer, 2019], on the other hand, assumes no other information than the allele frequency from pool sequencing (Pool-Seq) and aims to reconstruct selected haplotype blocks. This heuristic approach, however, infers only a small subset of SNPs on one of the haplotypes using allele frequency data from a sufficient number of replicate populations and generations. Nevertheless, it has been shown that even limited haplotype information is very helpful to infer selection and understand the genetic architecture of adaptation ([Michalak et al., 2019], [Mallard et al., 2018], [Karasov et al., 2010], [Barghi et al., 2019], [Burke, 2012]).

Here, we propose a new principled approach for a situation, where only allele frequency data are available, but no candidate haplotypes from other sources, as e.g. in [Griffin et al., 2017]. The approach builds on recent work by [Behr and Munk, 2017], and may be applied for instance with several Pool-Seq samples. Our focus is on samples collected over time, but data from multiple spatial locations would be another possible application. For the approach to work, several samples showing a sufficient fluctuation in haplotype frequencies are needed. For temporal data this condition is met when selection acts on the haplotypes, or when genetic drift is sufficiently large. For spatial data, samples from a sufficiently structured population would be needed. As the number of haplotypes that can be inferred reliably from allele frequency data is typically lower than the number of available samples, often only the most common haplotypes will be reconstructed.

Here we focus on time series from only one population. For experiments starting with a common pool of founder haplotypes, a simultaneous analysis of several replicate populations may lead to a larger number of sequenced samples and consequently to more accurate estimates from our method. Such a design also permits to address additional biologically interesting questions about the genetic redundancy of adaptation, by looking at the consistency in the haplotype frequency changes across replicates.

3 Methods

Notation In the following for an integer N we use the notation $[N] := \{1, \dots, N\}$. For a matrix A we let $A_{i\cdot}$ and $A_{\cdot i}$ denote its ith row and column vector. With $\|A\|$ and $A^\top$ we denote the Frobenius norm and the transpose of a matrix A, respectively. For a vector a, we always assume it is a column vector. We denote $\mathbf{1} = (1, \dots, 1)^\top$ the vector with just ones.

Simultaneously reconstructing the structure of dominant haplotypes (a.k.a. major haplotypes) as well as their relative proportions in the population amounts to a matrix factorization problem with finite alphabet constraint on one of the matrices, and positivity as well as unit column-sums constraints on the other. More specifically, assume we obtained relative allele frequencies $Y \in [0,1]^{N \times T}$ from a pool sequencing experiment, at time points $t \in [T]$ and SNP locations $n \in [N]$, from a population that consists of m_0 haplotypes. Then the underlying

population allele frequencies $F \in [0,1]^{N\times T}$ can be written as

$$F = S\,W, \qquad S \in \{0,1\}^{N\times m_0}, W \in [0,1]^{m_0\times T},$$

where $S_{\cdot i} \in \{0,1\}^N$ for $i \in [m_0]$ denotes the genotype structure of haplotype i, that is, $S_{ni} = 1$ if haplotype i takes the reference allele at location n and $S_{ni} = 0$ otherwise. The frequencies W_{it} denote the relative proportion of haplotype i at time point t (haplotype frequency). Ignoring any sequencing error, we have $E(Y|F) = F$. Our aim is to reconstruct both the matrices S and W from the measurement matrix Y. This amounts to a specific type of a *finite alphabet blind separation* problem [Behr and Munk, 2017, Behr et al., 2018].

In general, m_0, the overall number of haplotypes, can be very large, possibly $m_0 > n$, which makes S and W non-identifiable, even from the noiseless allele frequencies F. However, when (for most of the time points t) the population is dominated by $m \ll m_0$ haplotypes, such that,

$$\|W_{1\cdot}\| \geq \cdots \geq \|W_{m\cdot}\| \gg \|W_{(m+1)\cdot}\| \geq \cdots \geq \|W_{m_0\cdot}\|,$$

we denote structure and frequency of the dominant haplotyes as $S^d = (S_{ni})_{1\leq n\leq N, 1\leq i\leq m}$ and $W^d = (W_{it})_{1\leq i\leq m, 1\leq t\leq T}$ and obtain $F = S^dW^d + B$, with a bias $B = SW - S^dW^d$, which is the allele frequency component of minor haplotypes. In the following, we will omit the superscript d and just write

$$E(Y|F) = SW + B, \text{ with} \qquad S \in \{0,1\}^{N\times m}, W \in [0,1]^{m\times T}$$

such that $m \ll T, N$ and $\|B\| \ll \|SW_{i\cdot}\|$ for $i \in [m]$. With our considered simulation and real data scenarios, a bias term only makes a difference when there is a large number of minor haplotypes present at several time points. For our further analysis we assume that the minor haplotypes B and the major haplotypes SW are independent. Treating the bias term B in a Bayesian setting, we assume that B_{nt} is a random variable with mean

$$E(B_{nt}) = b_t \text{ for all } n \in [N], t \in [T]. \tag{1}$$

In total, we obtain that

$$E(Y|SW) = SW + \mathbf{1}b^\top, \tag{2}$$

where $b = (b_1, \ldots, b_T)^\top \in [0,1]^T$ is the bias term from the minor haplotype contribution.

3.1 HaploSep algorithm

If one directly had observed the allele frequencies $F = SW + \mathbf{1}b^\top$, one would be able to uniquely recover S, W and b by exploring the ordering structure of the rows of F (assuming some weak identifiability conditions on S and W as detailed in the SI). For example, the row vector $F_{i\cdot}$ with the smallest norm corresponds to a haplotype structure where $S_{i\cdot} = (0, \ldots, 0)$ and thus, $F_{i\cdot} = b$, which allows to recover b. Similar, the second smallest row vector of F corresponds to the situation where $S_{i\cdot} = (0, \ldots, 0, 1)$ and $F_{i\cdot} = W_{m\cdot} + b$, which allows to recover $W_{m\cdot}$. Proceeding in an analog way, one can uniquely recover S, W and b. Details are given in the SI and pseudo code is given in Algorithm 1 (SI).

However, in practice, one only obtains the noisy pool sequencing data Y but not the population allele frequencies F. Therefore, a direct application of Algorithm 1 (SI) is impossible and also not reasonable as further regulation will be required to obtain statistically stable estimates of S, W and b. Therefore, we consider a relaxation of the exact solution to $Y = SW + b$, i.e.

we seek to solve the optimization problem

$$\begin{aligned} &\hat{S}, \hat{W}, \hat{b} \in \arg\min_{S,W,b} \|Y - SW - \mathbf{1}b^\top\| \text{ with} \\ &S_{ni} \in \{0,1\},\ W_{it}, b_t \in [0,1],\ \sum_{i=1}^{m} W_{it} \leq 1 \end{aligned} \tag{3}$$

for $i \in [m]$, $t \in [T]$, $n \in [N]$. If Y were normally distributed, this would be the maximum likelihood estimator. For Pool-Seq data, due to its discrete structure, Y is clearly not normally distributed. Therefore, in principle, one may try to model the noise distribution of Y more precisely, and use a more targeted loss function than the L^2-norm loss in (3). However, in our simulations we found that loss functions based on a binomial model for the pool sequencing procedure do not provide a significant improvement over the L^2-norm loss and are computationally more challenging. This may be caused in part by the unpredictable variation of the bias term B.

Due to the discrete nature of S, the optimization problem in (3) is highly non-convex, which reveals this as a difficult issue. However, conditioned on either of (W, b) or S, optimization in (3) becomes tractable: Indeed, minimizing (3) given (W, b) corresponds to a simple clustering problem with known centers

$$\mathcal{C} = \{sW - \mathbf{1}b^\top \ : \ s \in \{0,1\}^m\}. \tag{4}$$

Given S, on the other hand, minimizing (3) corresponds to a simple linear regression problem with linear constraints on W, b. Thus, a very natural approach to tackle the minimization problem in (3) is to employ an iterative Lloyd's type algorithm [Lu and Zhou, 2016]. That is, to initialize either S or (W, b) and update iteratively until convergence of $\|Y - \hat{S}\hat{W} - \hat{b}\|$. Recently, [Lu and Zhou, 2016] showed that for a sub-Gaussian error distribution and appropriate initialization of either labels or clusters, Lloyd's algorithm converges to an exponentially small clustering error in $\log(N)$ iterations. For generic Lloyd's they also show that spectral clustering provides an appropriate initialization. Here, however, we cannot initialize the centers directly, but rather have to initialize the frequencies W and the bias term b, which indirectly determine centers via (4). For this, we propose to employ Algorithm 1 in the SI applied to the clustered observations $Y_{n\cdot}$. Pseudo-code is given in Algorithm 3 in the SI.

Similar as Algorithm 1 yields exact recover of W in the noiseless population case $Y = F$, it can be shown that Algorithm 3 yields stable recovery of W whenever the estimated centers $\hat{\mathcal{C}}$ are close enough to the noiseless centers $\mathcal{C}$ in (4). More precisely, this means that whenever $\max_{c\in\mathcal{C}} \min_{\hat{c}\in\hat{\mathcal{C}}} \|c - \hat{c}\| \leq \epsilon$, then it follows that $\max_{i\in[m]} \|W_{i\cdot} - \hat{W}_{i\cdot}\| \leq \epsilon$, for any sufficiently small ϵ, see [Behr et al., 2018, Behr, 2018] for details. The pseudo-code in Algorithm 2 in the SI summarizes our complete procedure for iteratively recovering the haplotype structure S and frequencies W from data Y, using Algorithm 3 as initialization. The only tuning parameter of our procedure in Algorithm 2 is the threshold δ used in our iteration stopping criterion. We found that $\delta = 0.001$ works well in practice with convergence in usually a couple of iterations.

Note that in practice, the number of major haplotypes m is not given and has to be estimated from the data Y. Cross validation and bootstrap methods could be employed for this purpose. In Section S1-4, we present a different approach which is (computationally) much simpler and based on singular value decomposition. As further information on the reliability of our estimates, we propose accuracy measures and explain their computation in Section S1-4.

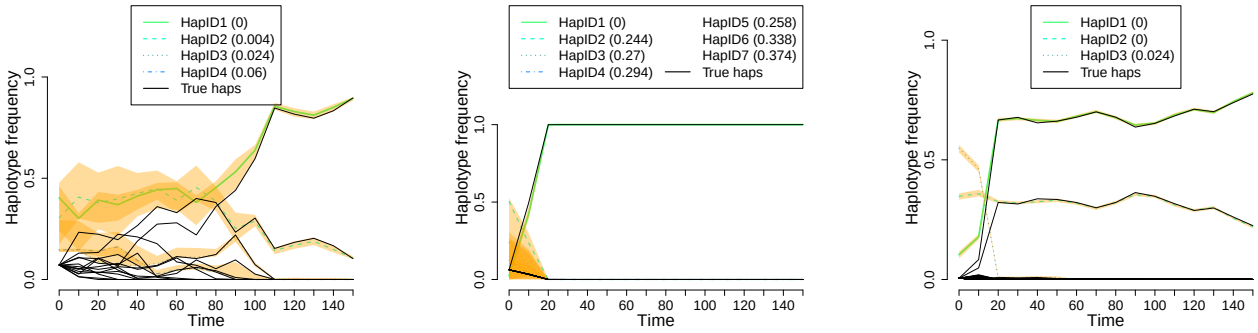
4 Simulations

Our simulations are designed to mimic experimental evolution (see e.g. [Kawecki et al., 2012], [Long et al., 2015] and [Schlötterer et al., 2015] for reviews). These experiments permit to

study evolutionary adaptation under controlled laboratory conditions, making it easier to disentangle adaptive responses from other factors such as demography or genetic drift. Typically multiple populations of organisms are kept in the laboratory for several generations under stressful conditions chosen by the experimenter. DNA sequence information is commonly obtained at different time points to study the genetic basis of adaptive responses. However the separate sequencing of individuals at high coverage will often be too time consuming or costly for larger populations. As a consequence, the analysis is frequently carried out based on estimated population allele frequencies from pools of individuals sequenced together, although haplotype information would be helpful for a better understanding of the adaptive process.
With our simulations, we intend to illustrate how haplotypes and their relative frequencies can be reconstructed in such experiments. Furthermore, we show that it is possible to obtain improved population allele frequency estimates with pool sequencing experiments by using the reconstructed haplotypes and their estimated frequencies.

4.1 Reconstruction of haplotype structure and frequency

To illustrate our method, we discuss a viability selection model with a selected locus (selection coefficient $s = 0.05$) occurring on three of the haplotypes. Drift is simulated every generation via multinomial sampling from the haplotypes present in the previous generation. We consider an experiment with a constant population size over 150 generations. To mimic pool sequencing, the allele frequency data are obtained via binomial sampling at a Poisson ($\lambda = 80$) coverage from the population allele frequencies. We chose starting haplotypes from a set of founder haplotypes sequenced by [Barghi et al., 2019]. Due to the shared genealogy, such haplotypes share a large amount of similarity, making the reconstruction challenging. In this context, we consider three scenarios that differ with respect to the population size and the number of starting haplotypes. These parameters were chosen to mimic the experiments described in Section 5, namely, experiments with *C. elegans*, *D. simulans*, and the Longshank mice experiment. Further details concerning the simulation setup can be found in Section S2 (SI).
To better understand the performance of our estimates, some exemplary situations are shown in Fig. 1 and the respective allele composition results in Fig. S1 (SI). Whereas the allelic composition of the major haplotypes is estimated with very low error in Fig. 1(a), the haplotype frequency estimates become accurate only at later stages of this experiment. In Fig. 1(b), on the other hand, the frequency estimates are very accurate except at the very beginning of the experiment, but only the composition of the dominating haplotype is accurately estimated. Finally, in Fig. 1(c) both frequencies and allelic composition of the dominating haplotypes are estimated very accurately. The explanation for these observed patterns is that several low frequency confounder haplotypes are present for a considerable time in the Longshank mice experiment, before most of them get eliminated by genetic drift. With our *C. elegans* example on the other hand, genetic drift eliminates all except one haplotype very quickly. The remaining haplotype is easily estimated, but for the disappearing ones, only very few time points provide information to reconstruct their composition. Therefore the reconstruction error is between 24% and 37%. Finally for *D. simulans*, there is sufficient information in the data to estimate the two dominating haplotypes more or less perfectly. Frequency and allelic composition is accurately inferred even for a third one.
For a more complete picture, we now report reconstruction errors for 100 simulation runs under the scenario mimicking the Longshank mice experiment in Fig. 2. The number of reconstructed haplotypes is estimated for each run via our model selection criterion explained in S1-3 (SI). The boxplots in (a) depict the errors in terms of the mismatch proportion for each of the reconstructed haplotypes, whereas (b) provides errors in terms of the mean absolute difference between the true and estimated frequencies at each time point. According to (a), the compo-

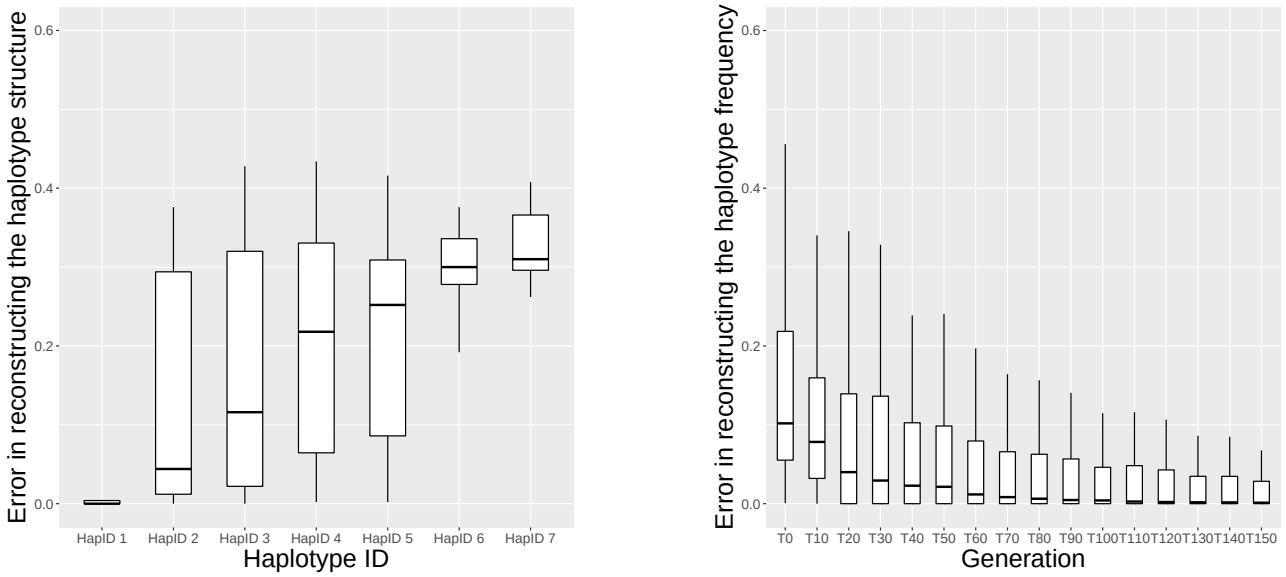


(a) *Longshank mice exp.* (b) *C. elegans* (c) *D. simulans*

Figure 1: Results of a typical simulation run using features of (a) the Longshank mice experiment, (b) the *C. elegans* design, and (c) the *D. simulans* experiment. True (black solid lines) and reconstructed haplotype frequencies are shown, with accuracy intervals (0.025 to 0.975 quantiles based on bootstrap) in yellow. In parentheses, we report the proportion of mismatches between estimated and the corresponding true haplotype. Details about the simulation scenarios can be found in the text and Section S2 (SI).

sition of the dominating haplotype (the one with the highest inferred frequency at the end of the experiment) is always estimated nearly perfectly. For the other haplotypes, the accuracy depends on whether the simulated trajectory provides enough information. As we simulated three of the haplotypes as selected, those haplotypes often (but not always) reached sufficiently high allele frequencies during the experiment and could therefore be estimated reliably. The frequency estimates in (b) clearly improve over time, illustrating again that accurate frequency estimates can be expected at time points where not too many haplotypes are present. Our simulations led to occasional outliers, i.e. situations where the accuracy is less satisfactory. For a practical application, we therefore recommend to use the accuracy scores R^2 proposed in Section S1-4 and the frequency change of reconstructed haplotypes for assessing the reliability of our estimates. In Fig. 2 for instance, we filtered out scenarios where either $R^2 < 0.8$ or the frequency change of dominating haplotype (HapID1) is below 0.1. These criteria generate a reasonable threshold to filter out a small proportion of problematic scenarios. For different experimental designs, we recommend to validate the thresholds with simulations. In supplementary Section S6, we provide a more detailed discussion of situations that may lead to outlying estimates.
For results simulated under the other two experimental setups, see Figs. S2 and S3 (SI).

When planning an experiment, it can be useful to know under which design parameters haplotype reconstruction will tend to be reliable. For this purpose, we provide simulation results exploring the influence of E&R designs on the accuracy of our method and summarize the results in the supplement. See Section S2 for a summary of the simulated scenarios, and Section S3 for the results obtained under these scenarios. While our simulations suggest a good performance over a wide range of scenarios, we recommend that potential users perform additional simulations, if their experimental design deviates from our considered scenarios.
A general observation is that the more the haplotypes change in their frequencies during the experiment, the better the haplotype reconstruction gets. In E&R this can be achieved through a large enough selection pressure affecting the investigated genomic region, or through small population sizes such that genetic drift causes large frequency changes. In other applications, where samples may differ in location rather than time, a sufficient amount of population structure would be needed.



(a) *Structure* (b) *Frequency*

Figure 2: Haplotype reconstruction error for our basic selection scenario with Longshank mice 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. For boxplots the function *geom_boxplot* from **ggplot2** (version 3.2.0) is used. The lower and upper hinges correspond to the first and third quartiles. The whiskers extend to 1.5 * IQR from the hinges (where IQR is the inter-quartile range). The same applies to all boxplots in the manuscript.

4.2 Improved allele frequency estimates

With known founder haplotypes, it has been shown in [Tilk et al., 2019] that allele frequency estimates from pool sequencing can often be improved by using haplotype information. Here we investigate, if this observation can be extended to the case of unknown founder haplotypes by using our estimates of important underlying haplotypes and their frequencies. Indeed, allele frequency estimates can be obtained by multiplying the matrix of the reconstructed haplotype structure ($\hat{S}$) with the matrix of the estimated haplotype frequencies ($\hat{W}$) and adding the estimated bias term $\hat{b}$. Using the simulated data in Section 4.1, we compared the so obtained estimates with the original allele frequencies from pool sequencing. As a measure of the difference in accuracy, we computed the ratio

$$\alpha = \frac{\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i^{\text{haplotypes}}|}{\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i^{\text{pool}}|}$$

where N is the number of SNPs, y_i is the true allele frequency of SNP i, $\hat{y}_i^{\text{haplotypes}}$ is the allele frequency of SNP i estimated using the reconstructed haplotypes, and $\hat{y}_i^{\text{pool}}$ is the one estimated by pool sequencing. If α is smaller than one, the haplotype based estimate performs better.

For each time point, we computed α based on all SNPs where the allele is not fixed or lost. To eliminate situations where the haplotype reconstruction does not work so well, we filter using the criterion in S1-4 to decide whether to use the haplotype based allele frequency estimates. Fig. 3 summarizes the relative performance for the Longshank mice experiment based on the filtered data. Analogous results for the other two experimental designs can be found in Fig. S14 (SI). Our results reveal that the use of the reconstructed haplotype information

typically leads to an improved accuracy. Indeed, since haplotype frequency estimates combine information across many SNPs, they are less noisy than allele frequencies from pool sequencing for individual SNPs.

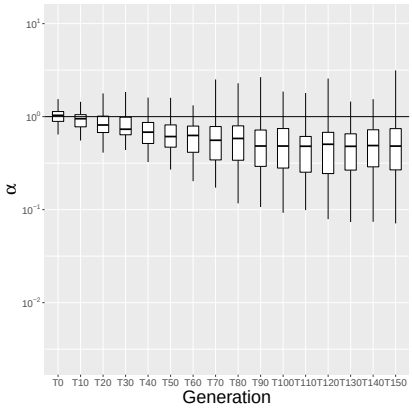


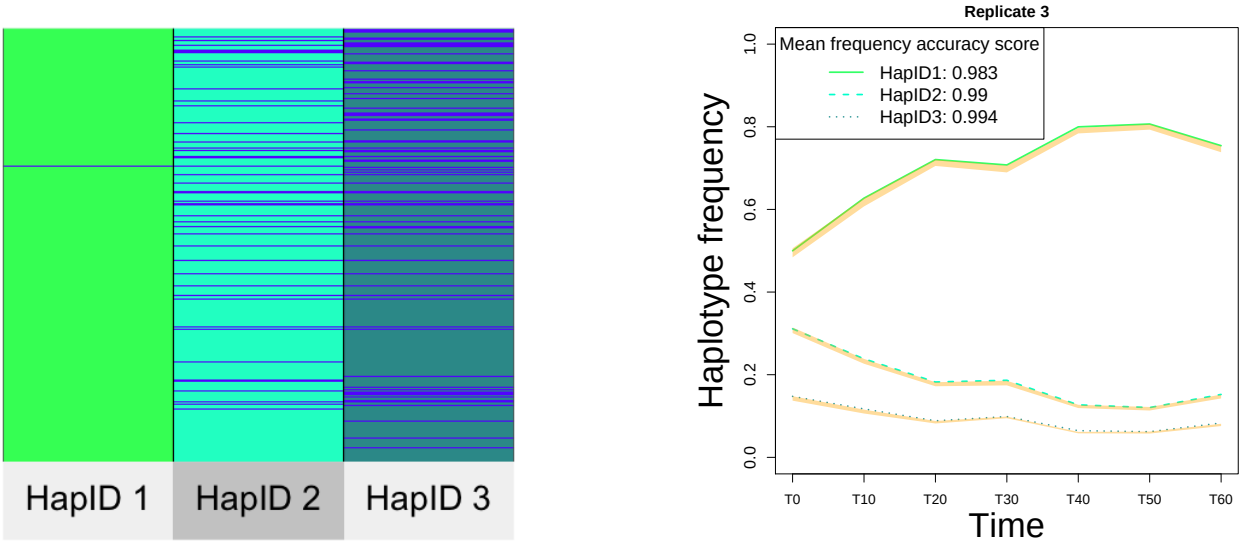
Figure 3: Error ratio (α) between haplotype based allele frequency estimates (numerator) and the pool sequencing estimates (denominator) plotted on a log-scale. Results from 100 simulations based on the Longshank mice experimental design. Except for the first time point ($T0$), the haplotype based estimates are usually more accurate.

5 Application to real data

We applied our approach to two E&R data sets taken from [Barghi et al., 2019], and [Noble et al., 2019], and to the not yet published data set described in [Castro et al., 2019]. In each case, we looked at a small genomic region under selection. With the data set from [Barghi et al., 2019] and [Noble et al., 2019], we also compared our inferred haplotypes with reference haplotypes provided by the authors.
As a further validation of our approach, we compared our reconstructed haplotypes with paired end reads using the original sequencing data from [Barghi et al., 2019]. We found our reconstructed haplotypes to be vastly concordant with the reads when interpreting them as very short haplotypes. For further details see Section S8 (SI).

5.1 *Drosophila simulans (D. simulans)*

We now consider the E&R experiment of [Barghi et al., 2019]. There the base population consists of 202 isofemale lines of *Drosophila simulans*. Ten replicate populations were kept for 60 generations, with sequencing data available every ten generations. Furthermore, a sample of 189 founder haplotypes was sequenced, as well as 100 additional ones from five evolved replicates.
For this data, we first identified interesting genomic regions by testing for signals of selection at the SNP level using the modified χ^2 and Cochran–Mantel–Haenszel tests [Spitzer et al., 2020] that account for drift and sequencing noise in the data. We then applied our method to multiple regions showing statistically significant allele frequency changes. We considered positions 11.239636 to 11.591566 Mb on chromosome 2L for replicate 3 as an example. Fig. 4 provides the estimated haplotype trajectories, as well as a comparison between our estimated allelic composition and the best matching founder haplotype sequences.
Due to the presence of a large number of similar haplotypes in particular at the early generations, the reconstruction is quite challenging for this experiment. Nevertheless, the dominating haplotype is usually reconstructed almost without error.



(a) *Structure* (b) *Frequency*

Figure 4: (a) Match between reconstructed haplotype structure and sequenced founder haplotypes for the region 11.239636 to 11.591566 Mb on chromosome 2L (replicate population 3) from the *Drosophila simulans* experiment. Blue lines indicate mismatches. This plot is generated using the superheat function from [Barter and Yu, 2018]. (b) Reconstructed haplotype frequencies with accuracy intervals (in yellow) and mean accuracy scores, with data as in (a). The displayed accuracy intervals for haplotype frequencies in (b) and throughout the paper are the 0.025 and 0.975 quantiles of $\hat{W}_{it}(Y^*)$ as detailed in S1-4.

5.2 Longshank experiment in mice

In the Longshank mice experiment, individuals from a mouse population were selected to produce offspring according to their tibia length/body mass ratio. The evolution of three populations (two Longshank lines and one control line) was followed over several generations. For details on the experimental design, we refer to [Castro et al., 2019] and [Marchini et al., 2014]. We received time series data from this experiment for the the Nkx3-2 region (4395 SNPs and indels) of the Longshank 1 (LS1) line collected every generation, from generation 0 to 20 (still unpublished). The allele frequencies were missing for a large number of SNPs. Therefore we decided to remove the later generations (14-20) from our analysis, because of their particularly high proportion of missing values. For the generations 0-13, we only kept those SNPs for which all allele frequencies were available. Filtering the data this way, we reconstructed haplotypes from the remaining 561 SNPs in this region.
The right panel of Fig. S22 displays our estimated haplotype trajectories and the corresponding accuracy scores. Unfortunately no founder haplotypes or read data for comparison purposes were available to us with this experiment.

5.3 *Caenorhabditis elegans (C. elegans)*

We next look at the experiment described in [Noble et al., 2019]. There, three replicate populations experienced an increasing quantity of NaCl during their evolution. The base population comprised 10^4 individuals that originated from 16 founder inbred lines. Pool sequenced allele frequency data have been made available to us for the base population and for the evolved populations at generations 50 and 100. Additionally, sequence information for the 16 founder inbred lines, as well as low coverage sequencing of a few individuals from the base population

and the two subsequent time points has been provided. Further details on the experimental design can be found in [Noble et al., 2017].
As for section 5.1 we searched for genomic regions showing signatures of selection. As an example, we provide results for a genomic region containing 666 SNPs (chromosome 5, 14924777-15216613 bp). The upper panel of Fig. S23 (SI) illustrates the close match between the reconstructed haplotypes and the most similar sequenced founder haplotype. For each replicate line, the lower panel of Fig. S23 (SI) shows the reconstructed haplotype trajectories, together with the corresponding accuracy measures. Moreover, since the three replicates show similar evolutionary patterns, we decided to also apply our method to all of them simultaneously. The result seems consistent with our single replicate analysis in terms of the reconstructed haplotypes and their frequency trajectories. This suggests parallel evolution across the replicates (see Fig. S24 in the SI).

6 Discussion

We proposed a new principled approach that for the first time estimates multiple completely unknown haplotypes from allele frequency data only. Under a suitably chosen experimental design, and with sufficiently large allele frequency changes, the allelic composition of the unknown haplotypes can be recovered reliably. Strong enough selection provides one scenario that leads to sufficient fluctuations in allele frequency.
A good reconstruction of haplotype frequencies is achieved at time when a moderate number of haplotypes is present at a sufficiently high frequency which may not be the case at early time points when an experiment starts with many founder haplotypes. Further important design parameters that affect the quality of our reconstruction are the population size, and the number of samples with sequence information.
A lot of scientific studies use haplotype information to answer their research questions. By providing estimates of the most important underlying haplotypes, our method will help researchers that only have allele frequency data available.
Our estimated haplotype frequencies can also be used to obtain allele frequency estimates that are less noisy on average than the original frequency data obtained for instance via pool sequencing. Indeed, by combining information from several neighboring SNPs, the sampling variation introduced by sequencing a whole pool of individuals gets averaged out to some extent.
As our next step, we plan to extend our approach to data from locally structured populations, where samples are usually taken from multiple subpopulations. The reconstructed haplotypes may also be helpful to impute missing data from low coverage sequencing. Another plan for subsequent research is to use information from the paired end read data directly with our estimates, as these reads might be interpreted as very short observed haplotypes.
We implemented our method *haploSep* in an *R* package available on Github at `https://github.com/MartaPelizzola/haploSep`.

Acknowledgement

We are grateful to the laboratories of Nick Barton, Christian Schlötterer, and Henrique Teotonio for providing us with their experimental data. This work has been supported by the Austrian Science Fund (FWF Doctoral Program Vienna Graduate School of Population Genetics", DK W1225-B20). MB was supported by Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) Postdoctoral Fellowship BE 6805/1-1. Moreover, MB acknowledges funding of DFG-GRK 2088. This work benefited from a research stay that was partially supported by the Simons Foundation and by the Mathematisches Forschungsinstitut Oberwol-

fach. AM and MB acknowledge support of DFG-SFB 803 Z02. AM and HL are funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2067/1 - 390729940.

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.

[Barter and Yu, 2018] Barter, R. L. and Yu, B. (2018). Superheat: An R Package for Creating Beautiful and Extendable Heatmaps for Visualizing Complex Data. *Journal of Computational and Graphical Statistics*, 27(4):910–922.

[Behr, 2018] Behr, M. (2018). *Finite Alphabet Blind Separation*. PhD thesis, University of Goettingen.

[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.

[Browning et al., 2018] Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics*, 103(3):338–348.

[Browning and Browning, 2011] Browning, S. R. and Browning, B. L. (2011). Haplotype phasing: Existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714.

[Burke, 2012] Burke, M. K. (2012). How does adaptation sweep through the genome? Insights from long-term selection experiments. *Proceedings of the Royal Society B: Biological Sciences*, 279(1749):5029–5038.

[Burke et al., 2010] 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.

[Cao and Sun, 2015] Cao, C.-C. and Sun, X. (2015). Accurate estimation of haplotype frequency from pooled sequencing data and cost-effective identification of rare haplotype carriers by overlapping pool sequencing. *Bioinformatics*, 31(4):515–522.

[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.

[Delaneau et al., 2019] Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L., and Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, 10(1):1–10.

[Excoffier and Slatkin, 1995] Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12(5):921–7.

[Franssen et al., 2017] Franssen, S. U., Barton, N. H., and Schlötterer, C. (2017). Reconstruction of Haplotype-Blocks Selected during Experimental Evolution. *Molecular Biology and Evolution*, 34(1):174–184.

[Futschik and Schlötterer, 2010] Futschik, A. and Schlötterer, C. (2010). The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics*, 186:207–18.

[Gasbarra et al., 2011] Gasbarra, D., Kulathinal, S., Pirinen, M., and Sillanpää, M. J. (2011). Estimating Haplotype Frequencies by Combining Data from Large DNA Pools with Database Information. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 8:36–44.

[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.

[Illingworth et al., 2012] Illingworth, C. J., Parts, L., Schiffels, S., Liti, G., and Mustonen, V. (2012). Quantifying selection acting on a complex trait using allele frequency time series data. *Molecular Biology and Evolution*, 29(4):1187–1197.

[Karasov et al., 2010] Karasov, T., Messer, P. W., and Petrov, D. A. (2010). Evidence that Adaptation in Drosophila Is Not Limited by Mutation at Single Sites. *PLoS Genet*, 6(6):e1000924.

[Kawecki et al., 2012] 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.

[Kessner et al., 2013] Kessner, D., Turner, T. L., and Novembre, J. (2013). Maximum likelihood estimation of frequencies of known haplotypes from pooled sequence data. *Molecular Biology and Evolution*, 30(5):1145–58.

[Loh et al., 2016] Loh, P. R., Palamara, P. F., and Price, A. L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics*, 48(7):811–816.

[Long et al., 2015] 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):567–82.

[Long et al., 2011] Long, Q., Jeffares, D. C., Zhang, Q., Ye, K., Nizhynska, V., Ning, Z., Tyler-Smith, C., and Nordborg, M. (2011). PoolHap: inferring haplotype frequencies from pooled samples by next generation sequencing. *PLoS One*, 6(1):e15292.

[Lu and Zhou, 2016] Lu, Y. and Zhou, H. H. (2016). Statistical and computational guarantees of Lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*.

[Mallard et al., 2018] Mallard, F., Nolte, V., Tobler, R., Kapun, M., and Schlötterer, C. (2018). A simple genetic basis of adaptation to a novel thermal environment results in complex metabolic rewiring in Drosophila. *Genome Biology*, 19(1):119.

[Marchini et al., 2007] Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7):906–913.

[Marchini et al., 2014] Marchini, M., Sparrow, L. M., Cosman, M. N., Dowhanik, A., Krueger, C. B., Hallgrimsson, B., and Rolian, C. (2014). Impacts of genetic correlation on the independent evolution of body mass and skeletal size in mammals. *BMC Evolutionary Biology*, 14(1):258.

[Michalak et al., 2019] 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.

[Noble et al., 2017] Noble, L. M., Chelo, I., Guzella, T., Afonso, B., Riccardi, D. D., Ammerman, P., Dayarian, A., Carvalho, S., Crist, A., Pino-Querido, A., Shraiman, B., Rockman, M. V., and Teotónio, H. (2017). Polygenicity and epistasis underlie fitness-proximal traits in the Caenorhabditis elegans multiparental experimental evolution (CeMEE) panel. *Genetics*, 207(4):1663–1685.

[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*.

[Otte and Schlötterer, 2019] Otte, K. A. and Schlötterer, C. (2019). A generalised approach to detect selected haplotype blocks in Evolve and Resequence experiments. *bioRxiv preprint 691659; doi: https://doi.org/10.1101/691659*.

[Pirinen, 2009] Pirinen, M. (2009). Estimating population haplotype frequencies from pooled SNP data using incomplete database information. *Bioinformatics*, 25(24):3296–3302.

[Sabeti et al., 2002] Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., and Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837.

[Schlötterer et al., 2015] 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.

[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.

[Tewhey et al., 2011] Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J., and Schork, N. J. (2011). The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215–223.

[Tilk et al., 2019] Tilk, S., Bergland, A., Goodman, A., Schmidt, P., Petrov, D., and Greenblum, S. (2019). Accurate Allele Frequencies from Ultra-low Coverage Pool-Seq Samples in Evolve-and-Resequence Experiments. *G3*, 9:4159–4168.

[Tishkoff et al., 1996] Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., Bonné-Tamir, B., Santachiara-Benerecetti, A. S., Moral, P., Krings, M., Pääbo, S., Watson, E., Risch, N., Jenkins, T., and Kidd, K. K. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science*, 271(5254):1380–1387.

[Turner et al., 2011] 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):e1001336.