# High accuracy haplotype-derived allele frequencies from ultra-low coverage pool-seq samples

Susanne Tilk[1], Alan Bergland[1,2], Aaron Goodman[1], Paul Schmidt[3], Dmitri Petrov[1], Sharon Greenblum[1]

[1] *Department of Biology, Stanford University, Stanford CA 94305*
[2] *Department of Biology, University of Virginia, Charlottesville VA 22904*
[3] *Department of Biology, University of Pennsylvania, Philadelphia PA 19104*

## Abstract

Evolve-and-resequence experiments leverage next-generation sequencing technology to track allele frequency dynamics of populations as they evolve. While previous work has shown that adaptive alleles can be detected by comparing frequency trajectories from many replicate populations, this power comes at the expense of high-coverage (>100x) sequencing of many pooled samples, which can be cost-prohibitive. Here we show that accurate estimates of allele frequencies can be achieved with very shallow sequencing depths (<5x) via inference of known founder haplotypes in small genomic windows. This technique can be used to efficiently estimate frequencies for any number of alleles in any model system. Using both experimentally-pooled and simulated samples of *Drosophila melanogaster*, we show that haplotype inference can improve allele frequency accuracy by orders of magnitude, and that high accuracy is maintained after up to 200 generations of recombination, even in the presence of missing data or incomplete founder knowledge. By reducing sequencing costs without sacrificing accuracy, our method enables analysis of samples from more timepoints and replicates, increasing the statistical power to detect adaptive alleles.

## Introduction

A major barrier to understanding the genetic basis of rapid adaptation has been the lack of robust experimental frameworks for assaying allele frequency dynamics. Recently, evolve and resequence (E+R) experiments[1], which leverage next-generation sequencing technology to track real-time genome-wide allele frequency changes during evolution, have become a powerful step forward in studying adaptation[2]. In most E+R studies, replicate populations are evolved over tens to hundreds of generations in an artificial or natural selection regime and allele frequency measurements from multiple timepoints are used to identify genomic targets of selection. To date, E+R approaches have already been successfully applied in a variety of model systems, including RNA molecules, viruses, *Escherichia coli*, *Saccharomyces cerevisiae*, and *Drosophila melanogaster*[3–7]. The ability to concurrently observe both trait and genome-wide allele frequency changes across multiple systems offers the potential to answer long-standing questions in molecular evolution. Careful analysis of the patterns and magnitude of allele frequency change may reveal the extent of the genome that is under selection, how interacting alleles contribute to adaptive traits, and the speed of adaptation in different evolutionary regimes.

Crucially, however, the power to address such questions depends on the replication, time-resolution, and accuracy of allele frequency trajectories, with tradeoffs between these often incurred due to high sequencing costs. Recommended E+R schemes with even minimal power to detect selection involve sampling tens to hundreds of individuals from at least three replicate populations over a minimum of ten generations[8,9]. Since individual-based, genome-wide DNA sequencing at sufficient coverages is enormously cost-prohibitive, most E+R studies rely instead on pooled sequencing[10–13] of all individuals sampled from a given timepoint and replicate. While this approach sacrifices information about individual genotypes and linkage, pooled sequencing has been shown to provide a reliable measure of population-level allele frequencies[14,15]. Still, forward-in-time simulations suggest that each pooled sample must be sequenced at a minimum of 50x coverage to detect strong selection and even higher coverage to detect weak selection[9]. Given that optimized experimental designs often involve >100 samples, total costs for *Drosophila melanogaster* E+R experiments that achieve reasonable detection power can reach well above $25,000 at current sequencing costs. Thus achieving sufficient accuracy remains a major limiting factor in capitalizing on the promise of E+R.

The short timescales for which E+R is most appropriate may, however, facilitate ways to reduce sequencing costs without sacrificing experimental power. First, there is a growing body of evidence that in sexual populations, the bulk of short-term adaptation, especially in fairly small populations, relies on standing genetic variation rather than new mutations[5,12]. Many E+R schemes involve experimental populations derived from a fixed number of inbred founder lines[6,16,17], so the identity, starting frequency, and haplotype structure of all segregating variants are often either already well-known or can easily be obtained by sequencing each founder line. Tracking only the frequencies of these validated variants can still provide enough power to detect selection, while reducing the depth of sequencing usually required to differentiate new mutations from sequencing error.

Second, at short timescales haplotype structure can be leveraged to provide more accurate allele frequency estimates. In the time frame of most E+R experiments, recombination does not fully break apart haplotype blocks and disrupt linkage, and thus genomes in an evolving population are each expected to be a mosaic of founder haplotypes. In this scenario, recently developed haplotype inference tools[18–23] can integrate information from sequencing reads across multiple nearby sites to efficiently infer the relative frequency of each founder haplotype within small genomic windows. These haplotype frequencies can then be used as weights to calculate pooled allele frequencies for local segregating variants. With this approach, the accuracy of an allele frequency estimate depends less on the number of mapped reads at the individual site, and instead relies on the discriminatory power of all mapped reads in the surrounding genomic window when inferring haplotype frequencies. Haplotype inference methods such as HARP[22] have been shown to accurately predict haplotype frequencies at coverages as low as 25x, and simulations of pool-seq data from a small genomic region at fixed read depth indicate that the use of haplotype frequency information increases the power to detect selection[24]. However, these tools have not yet been used to infer allele frequencies from real pooled samples in an E+R framework, nor has a thorough analysis been performed to fully examine how empirical depth of pooled coverage scales with the accuracy of haplotype-derived allele frequency estimates at individual SNPs genome-wide, across many parameters relevant for E+R.

Here, we focus on defining the accuracy of haplotype-derived allele frequencies (HAFs) in order to provide new optimal design recommendations for haplotype-informed E+R given realistic financial constraints, real-life experimental noise, and real-life levels of missing data. Since haplotype inference will be affected by 1) read depths throughout genomic windows, 2) recombination events, and 3) incomplete founder information, we begin by leveraging both simulated and experimental data to assess how the accuracy of HAFs scales with each of these parameters. To do so, we introduce a new metric, ‘effective coverage’, that associates the error from HAF estimates to the expected sampling error of pooled sequences at various read depths. We find that haplotype inference can significantly increase the accuracy of allele frequency estimations across a range of genomic window sizes, multiple generations of recombination, and with incomplete information about founders. Although we primarily focus on simulated and experimental data from *Drosophila melanogaster*, we later describe how our results can be extended to other model organisms as well. We conclude our findings by offering recommendations about the most powerful way to integrate haplotype inference into E+R experimental schemes, paving the way for deeper insight into the genomic underpinnings of adaptation.

## Results

Similar to many E+R studies that approximate population-wide allele frequencies at a given time-point by randomly sampling and pooling ~100 individuals[6,17,25], we pooled two biological replicates of 99 *Drosophila melanogaster* individuals, and performed high-coverage sequencing of each replicate. In our pooled samples however, each individual was derived from a different previously sequenced isogenic founder line.

All reads were mapped to the *D. melanogaster* reference genome, and non-HAFs and HAFs were calculated at each of the 283,437 known segregating bi-allelic sites on chromosome 2L (for simplicity, the remainder of the analysis focuses just on this chromosome). Non-HAFs were calculated by evaluating the fraction of mapped reads containing the alternate allele. To calculate HAFs, founder haplotypes were first constructed from founder genotype calls at the same 283k sites. Haplotype frequency estimation was performed with HARP, a haplotype inference tool that uses both sequence identity and base quality scores to probabilistically assign pool-seq reads to founder haplotypes, and then obtain maximum likelihood estimates of haplotype frequency in discrete chromosomal windows. After inferring the frequency of each founder haplotype in sliding windows across the chromosome, we calculated HAFs at each SNP site by evaluating the average weighted sum of local founder haplotypes containing the alternate allele.

To determine the accuracy of HAFs and non-HAFs, estimated allele frequencies were compared to ‘true’ allele frequencies derived from the known composition of founder haplotypes that incorporated estimates of uneven pooling (see Supp. Fig 2 and Supp. Text). Chromosome-wide accuracy of HAFs and non-HAFs was quantified using a new metric, *effective coverage,* which represents the theoretical coverage at which the expected binomial variance from read sampling equals the average error from observed allele frequency estimates (see *Methods* for full description). Note that while this metric specifically focuses on the variance from sampling of pooled sequences, in practice, the ability of both HAFs and non-HAFs to accurately reflect true population-level allele frequencies will also depend on variance from sampling of individuals from the population for pooling. This independent source of error has however been well-treated elsewhere[26,27] and will not be impacted by haplotype inference.

In the following sections, we explore how the accuracy of HAFs differs from non-HAFs, and how it scales with empirical coverage, inference window size, different founder sets, incomplete founder information, and the number of generations of recombination.

### *Imputation of missing genotype calls reduces bias in haplotype frequency assignment*

Ambiguous or missing founder genotype calls are common due to residual heterozygosity in inbred lines and uneven sampling during initial founder sequencing. On average, at each segregating site in our set of 99 founders, genotype calls for 4 founders were unresolved. Initial rounds of haplotype inference produced a clear negative correlation between the number of missing calls per founder and the haplotype frequencies estimated for that founder (Supp. Fig 1). However, we found that imputation of missing founder genotypes prior to haplotype inference both significantly reduced the skewed haplotype frequency assignment and produced more accurate HAFs overall. Thus, we include imputation as a key step in our HAF calculation pipeline for the rest of the analysis (see Supp. Text).

### *Haplotype inference significantly increases the accuracy of allele frequency estimations*

The accuracy of HAFs depends on the power to estimate haplotype frequencies, which in turn is affected by the coverage of mapped reads throughout the genomic window used for haplotype inference. In order to compare the accuracy of HAFs to non-HAFs and test how each scales with empirical coverage, reads from the two biological replicates originally sequenced at ~140x were down-sampled to chromosome-wide empirical coverages of 1x to 100x, and then used to calculate the effective coverage of allele frequencies for each replicate (Fig 1a). Haplotype inference was initially performed using 100kb sliding windows, and accuracy here is assessed at the 27k sites with known genotype information for every founder. As expected, effective coverage of both HAFs and non-HAFs is similar between the two biological replicates, and increases with greater chromosome-wide empirical coverage. Yet for all empirical coverages tested, HAFs have strikingly higher effective coverages than non-HAFs. This substantial gain in accuracy from haplotype inference was most prominent at lower empirical coverages, with a 40-fold increase in accuracy at 10x empirical coverage, from 10x to effectively 400x. At higher empirical coverages, haplotype inference appears to produce diminishing returns and effective coverage begins to plateau.

We next tested the effect of using smaller (10kb; Fig 1b) or larger (1000kb; Fig 1d) windows for haplotype inference at empirical coverages up to 10x. We find that for both pooled replicates, larger window sizes provide the most accurate HAFs, since more reads are available to infer haplotype frequencies in each window. Specifically, 1000kb HAFs derived from empirical coverages of 1x and 5x reached effective coverages of >400x and >900x, respectively.

We also confirmed that the same results would be achieved by simulated samples with known sources of error.  To do so, we simulated pooled synthetic reads with a standard Illumina sequencing error rate of 0.002[28] and corresponding base quality scores[22] from the same proportions of the 99 founder lines included in the first biological replicate, and calculated effective coverage with the same empirical coverages and window sizes as above. Effective coverages for these simulated samples closely mirror effective coverages obtained from matched experimental samples (Fig 1b-d). Slight differences at higher empirical coverages and larger window sizes are most likely caused by compounded experimental error from DNA extractions, PCR reactions and sequencing, as well as ambiguity in the ‘true’ genotypes estimated for individually sequenced lines.

Finally, we confirmed that the increased accuracy due to haplotype inference applies equally to pooled samples derived from other founder sets. We simulated pooled samples from an entirely different founder set composed of lines from the DGRP[29] and found that the relationship between HAFs and non-HAFs in all ranges of empirical coverages and window sizes tested is qualitatively the same between the two different founder sets (Supp. Fig. 3).

Together, these results suggest that HAFs derived from both simulated samples and multiple biologically distinct samples sequenced at low empirical coverages can be orders of magnitude more accurate than non-HAFs.  In the following analyses we focus on simulated data from the 99 original founder lines in order to precisely and reliably test how recombination and founder ambiguity affect HAFs in realistic E+R scenarios.

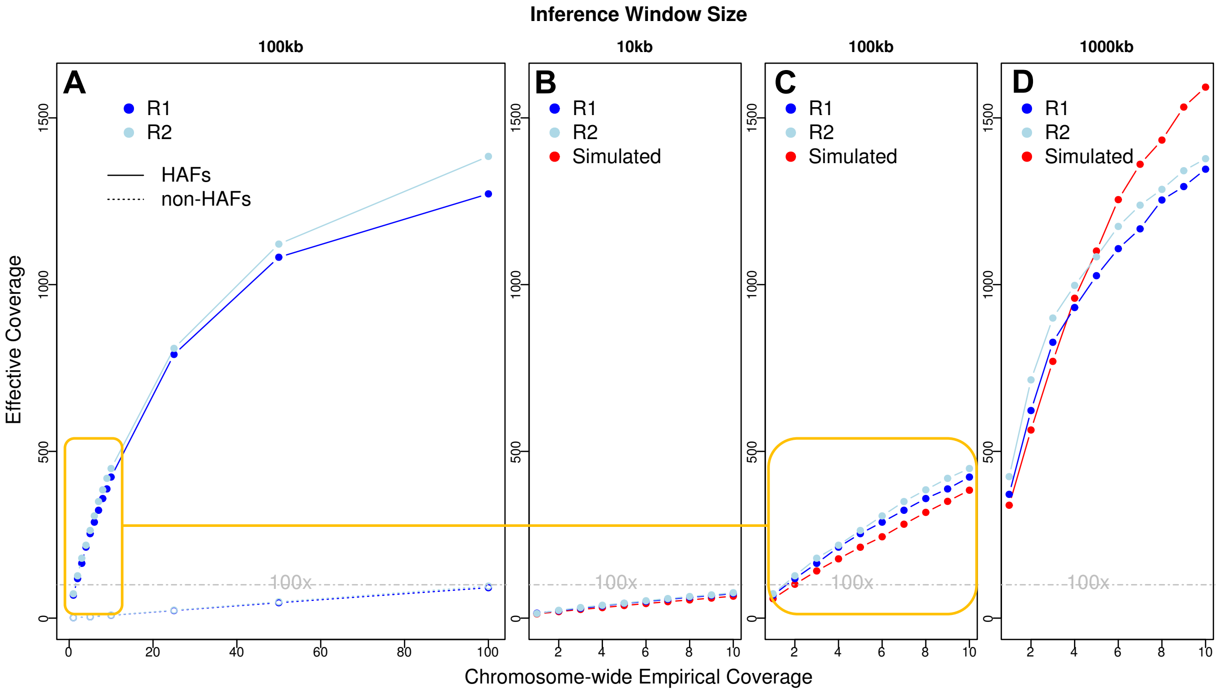


**Fig. 1. Accuracy of HAFs and non-HAFs for biological and simulated samples. A)** Effective coverage of allele frequencies estimated with and without haplotype inference (HAFs and non-HAFs, respectively) for two biological replicates of 99 pooled individuals sequenced to high coverage and down-sampled to empirical coverages from 1-100x (R1=replicate 1, R2=replicate; HAFs calculated with 100kb inference windows). **B-D)** Effective coverages of HAFs for biological replicates (blue) and simulated samples (red) using 10kb, 100kb, or 1000kb inference windows at empirical coverages of 1-10x.

### *Recombination and selection minimally affect accuracy over short time scales*

Since recombination will unlink neighboring alleles resulting in smaller haplotype blocks, we hypothesized that as an experiment proceeds, the accuracy of HAFs calculated with a given inference window size will decrease. To test this prediction, we used forqs[REF] to perform forward-in-time simulation of neutral recombination for 200 generations in a population of 1000 individuals using a *D. melanogaster* recombination map[30.], and tracked the breakpoints and haplotype of origin for all recombined segments at every generation. At various timepoints, we randomly selected 198 recombined chromosomes (ie. 99 diploid individuals), and constructed the full sequences of these 'sampled' chromosomes from corresponding segments of the 99 individually sequenced founder haplotypes. Reads were simulated from the pooled set of 198 chromosomes after assigning 'true' genotypes to all missing calls that were then hidden during downstream haplotype inference calculations (see methods for full description). The accuracy of HAFs calculated from these simulated reads was assessed using 1000kb, 100kb, and

10kb inference windows at all 283k segregating sites. This is contrast to the analysis above that only computed HAF accuracy at sites with complete founder information, and allowed us to realistically incorporate missing information found in experimental samples.

We first confirmed that, with all window sizes tested, the accuracy of HAFs decreases after more generations of recombination (Fig 2). Additionally, there is a tradeoff between window sizes (1000kb vs 100kb) after 50 generations, as the accuracy gained from extra information in larger windows is outweighed by the incorrect assumption of complete haplotype blocks. However, even after 50 generations of recombination, HAFs calculated with 100kb windows and an empirical coverage of 5x achieve an effective coverage of >100x, and the same empirical coverage can achieve effective coverages >50x at any timepoint tested. Conversely, 10kb windows may only be useful after >200 generations of recombination, longer than most E+R experiments to date, and therefore the rest of our analysis focuses on using 100kb and 1000kb windows.

In order to ensure that these results also extend to different evolutionary dynamics and are not specific any particular set of chromosomal breakpoints, we repeated this simulation 10 additional times, starting with the same founder population (Supp. Fig 4). During the first few generations of recombination, evolutionary dynamics and sampling do result in notable variance in the accuracy of HAFs, although this variance rapidly diminishes as recombination proceeds. We also confirmed that our estimates of effective coverage are robust to non-neutral dynamics in the presence of selected sites with varying selective strength (Supp. Fig 5).

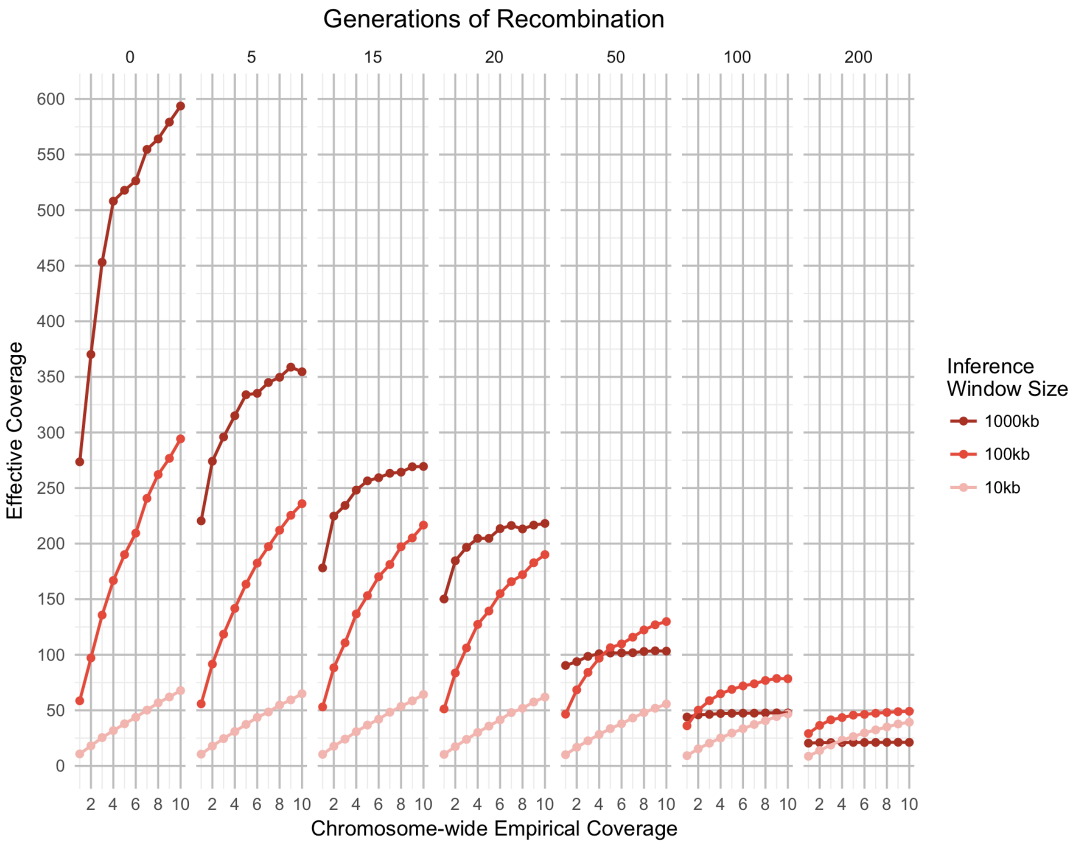


**Fig. 2. Effect of recombination on the accuracy of HAFs.** Effective coverage values of HAFs on chromosome 2L after 0 to 200 generations of recombination in simulated pooled samples of founders (n=99) using 1000kb (dark red), 100kb (red) and 10kb (pink) windows.

### *HAF accuracy is impacted by haplotype inference with incomplete founder sets*

Although recombination is a major factor affecting the accuracy of allele frequency estimates, ambiguity in the set of founders represented in the pool can also influence accuracy. As evolution proceeds during E+R experiments, whole founder haplotypes can be lost from the population. Similarly, contamination may create scenarios in which unintended haplotypes become part of the pool. Loss or gain of founder haplotypes over time has the potential to affect the accuracy of HAFs. We tested the accuracy of HAFs in the first scenario of haplotype loss by allowing haplotype frequency assignment to an extended set of founders, in addition to those included in the original population. Conversely, we tested the accuracy of HAFs in the second scenario by constraining haplotype frequency assignment to only 50% of the original founders. In both cases, HAF accuracy was evaluated in multiple window sizes after 5-20 generations of recombination.

We find that the first scenario (founder haplotype loss) has a notable effect on effective coverage (Fig. 3), reducing accuracy by ~40% and ~50% when 25% and 50% of the founders were lost from the population, respectively, after 15 generations of recombination at 10x coverage. The second scenario (haplotype contamination) resulted in a 50% decrease of effective coverage after 15 generations of recombination when the population was contaminated by additional haplotypes, increasing the number of haplotypes in the population by 25% compared to the original founder set of 99 haplotypes. Effective coverage was further decreased by 63% when the number of haplotypes in the population was increased by 50%, and by 77% when the number of haplotypes was doubled. However, even after founder losses and gains of 50%, effective coverage remained >100x and >50x, respectively, in all E+R scenarios tested here.

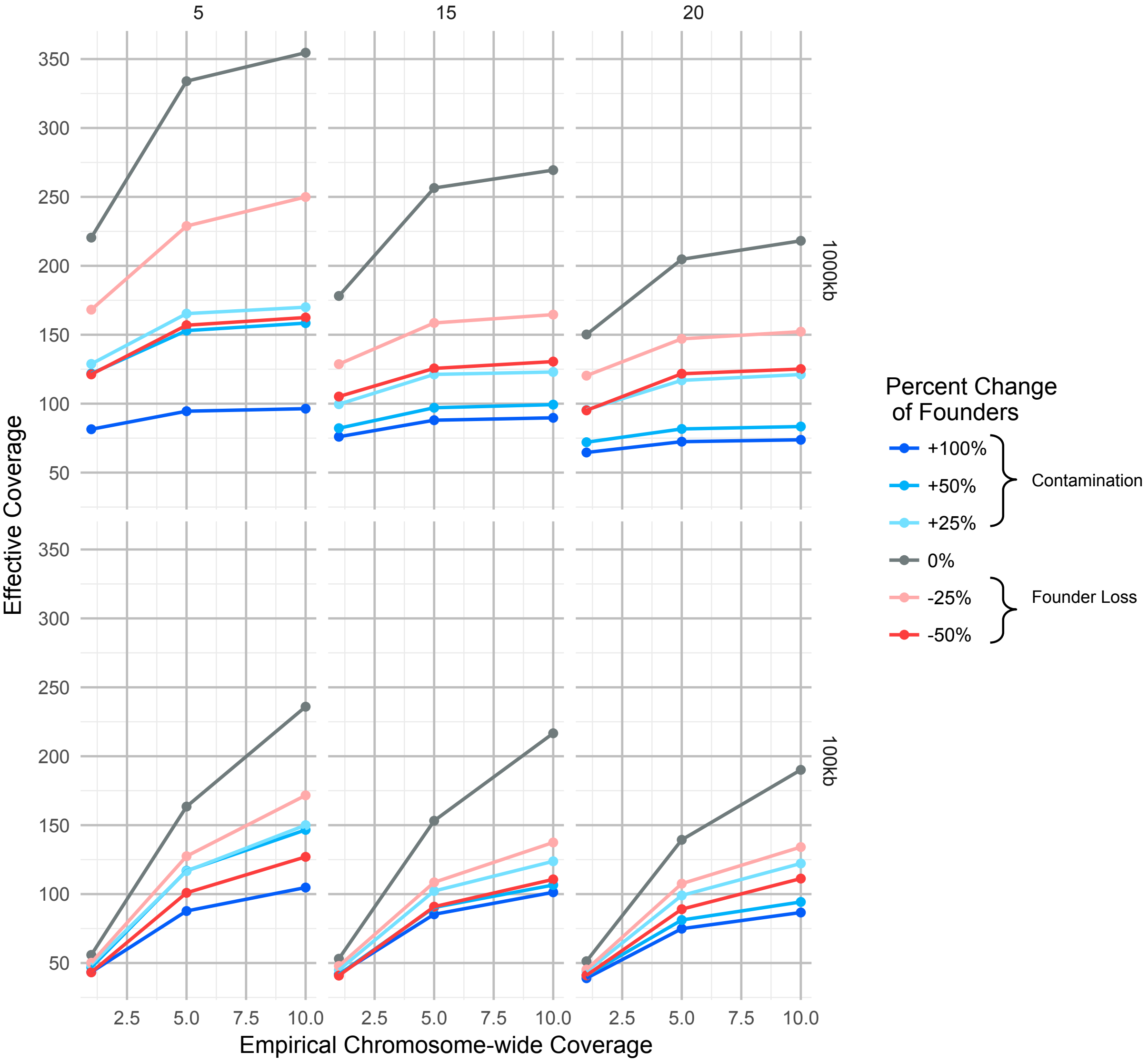


**Fig. 3. Measuring HAF accuracy after loss or gain of founder haplotypes.** Effective coverage values of HAFs after intermediate amounts of recombination (5 gens, left; 15 gens, middle; 20 gens, right) using 1000kb (top panel), or 100kb (bottom panel) windows. Effective coverage values were calculated after founder haplotype loss, where fewer founders were represented in the pool than the number of potential haplotype frequency assignments (red lines), and founder haplotype gain, where more founders were represented in the pool than the number of potential haplotype frequency assignments (blue lines).

### *Estimating effective coverage with other model organisms*

Finally, we explored how the utility of HAFs may extend to any genomic region, for any founder set with known SNPs, and in any model organism with a known recombination rate. To do this, we suggest using the chromosome-wide results presented in the analyses above as effective coverages expected at different empirical coverages and generations in *Drosophila* experiments, with the expectation that these values will vary slightly across the genome. The values reported here can be translated to any founder set or model organism by comparing SNP densities and recombination rates. For example, *Drosophila melanogaster* chromosome 2L has an average recombination rate of 2.39 cM/Mb[30] and our founder set contains an average of 27,050 SNPs per Mb. Chromosome I in *Caenorhabditis elegans*, however, has a recombination rate that is 1.4x that of Drosophila (3.33 cM/Mb[31]) and a commonly used reference panel of 249 *C. elegans* strains[32] contains only ~0.7x as many SNPs (18,600 SNPs/Mb). Thus, to achieve an effective coverage of 50x (the minimum required to detect selection) after 70 generations of *C. elegans* recombination, one should aim to sequence pooled samples at an empirical coverage of 7x.

## Discussion

E+R experiments have become a powerful tool to assay the underpinnings of rapid adaptation by tracking allele frequency trajectories within populations over time. Previous studies have shown that the greatest power to detect adaptive variants comes from an optimized experimental design that tracks allele frequencies in multiple replicate populations, samples each replicate population at multiple timepoints, and maximizes the coverage of each pooled sample. Incorporating all of these factors into an E+R framework, however, can present significant financial challenges. Here, we offer a way to mitigate these high sequencing costs without sacrificing statistical power.

Our framework uses haplotype inference to increase the accuracy of pooled allele frequency estimates at low coverages. Since the accuracy of haplotype-derived allele frequencies relies on the total discriminatory power of reads across a genomic window, rather than coverage at a single site, this approach allows us to sequence less but still maintain high accuracy in allele frequency estimations. Namely, our method achieves the same accuracy expected from sequencing each sample at 100x (as recommended in order to reliably detect strong selection), while only requiring empirical coverage of 1x or less, bringing total sequencing costs from >$25,000 down to less than $200.

There are, however, limitations to this approach. First, this framework requires the founder population to be derived entirely from fully inbred lines. As a result, the population dynamics of loci under selection may differ slightly from trajectories in natural populations due to the genetic diversity lost in the inbreeding process (i.e. natural haplotypes, homozygous lethal mutations, and rare variants), as well as higher levels of linkage disequilibrium. Reconstituting an outbred population using inbred lines, however, can be an effective way to mitigate effects of the inbreeding process, and has been experimentally shown to have negligible bias and effect on adaptive dynamics[33].

Second, this approach requires a reliable and comprehensive account of the variants present in each founder line. Since previous studies recommend upwards of 100 founders, sequencing each individual founder line to a sufficiently high depth may present a high upfront cost. However, this cost represents a one-time investment, which can be applied toward all future experiments using the same set of founders. Furthermore, a number of consortiums already maintain publically available stocks of large numbers of *Drosophila* lines with full, high-quality genome sequences[29,34]. We anticipate that these resources will continue to rapidly expand, facilitating experiments with even greater haplotype diversity at minimal costs.

In addition, this approach is limited to studying short-term adaptation on the scale of tens of generations. In fact, an assumption of our method is that within an inference window, recombination breakpoints minimally affect the ability to accurately call haplotype frequencies. For a given window size however, this assumption becomes less valid as recombination proceeds, and haplotypes blocks decay. Conversely, using increasingly smaller windows reduces the information used for haplotype inference, to the point at which HAFs are no longer more accurate than non-HAFs. Though our results here demonstrate that recombination will limit the ability to detect adaptation on timescales of more than 200 generations, the short-term adaptive dynamics which E+R is best suited for fall well within this range. Furthermore, it is at these short timescales, when large numbers of replicate populations are critical for reliably detecting selection, that the cost savings associated with haplotype inference methods will be most beneficial.

Finally, this approach relies on tracking the trajectories of known bi-allelic polymorphisms derived from the founder population, and thus, de novo mutations will not be assayed in this framework. Nonetheless our approach should sufficiently capture the salient features of short-term adaptive dynamics, as there is a growing body of experimental evidence suggesting that selection acts primarily on standing genetic variation in sexual organisms, and that de novo beneficial mutations do not play a large role in rapid adaptation[5,35–37]. Additionally, by tracking only known well-validated polymorphisms, the approach is largely robust to error from small non-SNP chromosomal variants such as indels.

Despite the above limitations, collectively our results show that integrating haplotype inference into future E+R experiments is the most cost-effective way to achieve accuracy in allele frequency estimates, which will directly improve the ability to detect genome-wide signatures of adaptation. Consequently, we offer specific recommendations for future E+R experimental schemes that take advantage of this approach. First, each founder line should be initially sequenced to a sufficient depth that minimizes any missing genotypes. Preliminary analysis reveals that the loss of even 1 called founder genotype per site results in notable drops in accuracy. If missing genotype calls do exist in founder lines, imputing sites prior to haplotype inference can mitigate some of this error. When calculating haplotype frequencies, we find that using large inference window sizes (1000kb) and providing information for the most comprehensive set of founders maximizes the accuracy of allele frequencies and effective coverages attained for each relevant E+R scenario tested here.

Together, these guidelines and the analysis above form a framework for achieving effective coverages of close to 100x with empirical coverages as low as 1x even after 50 generations of recombination, reducing sequencing costs by 100-fold. Ultimately, these cost savings, which can be extended to experiments with a variety of model organisms, will provide a more robust E+R framework that can incorporate large numbers of replicate populations. This will be crucial to the future of E+R as a sustainable and feasible experimental tool since it can provide the statistical power to distinguish between beneficial and neutral alleles.

## References

1. Long, A., Liti, G., Luptak, A. & Tenaillon, O. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nat Rev Genetics* **16,** 567–82 (2015).
2. Burke, MK. How does adaptation sweep through the genome? Insights from long-term selection experiments. *Proceedings of the Royal Society of …* (2012). at <http://rspb.royalsocietypublishing.org/content/early/2012/07/17/rspb.2012.0799.short>
3. Pitt, JN & Ferré-D'Amaré, AR. Rapid construction of empirical RNA fitness landscapes. *Science* (2010). at <http://science.sciencemag.org/content/330/6002/376.short>
4. Barrick, JE, Yu, DS, Yoon, SH, Jeong, H & Oh, TK. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* (2009). at <http://search.proquest.com/openview/6675f97b6f9c1df27431787caadadc8e/1?pq-origsite=gscholar&cbl=40569>
5. Burke, MK, Liti, G & Long, AD. Standing genetic variation drives repeatable experimental evolution in outcrossing populations of Saccharomyces cerevisiae. *Molecular biology and evolution* (2014). doi:10.1093/molbev/msu256
6. OROZCO-terWENGEL, P, Kapun, M & Molecular …, N.-V. Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Molecular …* (2012). doi:10.1111/j.1365-294X.2012.05673.x
7. Wichman, HA, Badgett, MR & Scott, LA. Different trajectories of parallel evolution during viral adaptation. … (1999). at <http://science.sciencemag.org/content/285/5426/422.short>
8. Kofler, R & Schlötterer, C. A guide for the design of evolve and resequencing studies. *Molecular biology and evolution* (2013). at <https://academic.oup.com/mbe/article-abstract/31/2/474/1000475>
9. Schlötterer, C, Kofler, R, Versace, E & Tobler, R. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. *Heredity* (2015). at <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4815507/>
10. Burke, MK, Dunham, JP, Shahrestani, P & Thornton, KR. Genome-wide analysis of a long-term evolution experiment with Drosophila. *Nature* (2010). at <http://search.proquest.com/openview/33de3b0d4f540e9392f9b5faff2a3368/1?pq-

origsite=gscholar&cbl=40569>

11. Illingworth, C. J., Parts, L., Schiffels, S., Liti, G. & Mustonen, V. Quantifying selection acting on a complex trait using allele frequency time series data. *Molecular biology and evolution* **29,** 1187–1197 (2011).

12. Graves, J. L. *et al.* Genomics of Parallel Experimental Evolution in Drosophila. *Molecular Biology and Evolution* doi:10.1093/molbev/msw282

13. Barghi, N., Tobler, R., Nolte, V. & Schlötterer, C. Drosophila simulans : A Species with Improved Resolution in Evolve and Resequence Studies. *G3 Amp 58 Genes Genomes Genetics* **7,** 2337–2343 (2017).

14. Zhu, Y, Bergland, AO, González, J & one, P.-D. Empirical validation of pooled whole genome population re-sequencing in Drosophila melanogaster. *PloS one* (2012). at <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0041901>

15. Fracassetti, M., Griffin, P. & Willi, Y. Validation of Pooled Whole-Genome Re-Sequencing in Arabidopsis lyrata. *Plos One* **10,** e0140462 (2015).

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

17. Jha, AR, Miles, CM, Lippert, NR & Brown, CD. Whole-genome resequencing of experimental populations reveals polygenic basis of egg-size variation in Drosophila melanogaster. *… biology and evolution* (2015). at <http://mbe.oxfordjournals.org/content/32/10/2616.short>

18. Cao, CC & Sun, X. Accurate estimation of haplotype frequency from pooled sequencing data and cost-effective identification of rare haplotype carriers by overlapping pool sequencing. *Bioinformatics* (2014). at <https://academic.oup.com/bioinformatics/article-abstract/31/4/515/2748165>

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

20. Pirinen, M. Estimating population haplotype frequencies from pooled SNP data using incomplete database information. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp584

21. Jajamovich, G. H., Iliadis, A., Anastassiou, D. & Wang, X. Maximum-parsimony haplotype frequencies inference based on a joint constrained sparse representation of pooled DNA. *BMC Bioinformatics* **14,** 270 (2013).

22. Kessner, D., Turner, T. & Novembre, J. Maximum Likelihood Estimation of Frequencies of Known Haplotypes from Pooled Sequence Data. *Molecular Biology and Evolution* **30,** (2013).

23. Franssen, S., Barton, N. & Schlötterer, C. Reconstruction of haplotype-blocks selected during experimental evolution. *Mol Biol Evol* **34,** 174–184 (2016).

24. Lynch, M, Bost, D, Wilson, S & biology and …, M.-T. Population-genetic inference from pooled-sequencing data. *Genome biology and …* (2014). at <https://academic.oup.com/gbe/article-abstract/6/5/1210/604081>

25. Turner, TL, Stewart, AD, Fields, AT & Rice, WR. Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in Drosophila melanogaster. *PLoS …* (2011). doi:10.1371/journal.pgen.1001336

26. Feder, A., Petrov, D. & Bergland, A. LDx: Estimation of Linkage Disequilibrium from High-Throughput Pooled Resequencing Data. *PLoS ONE* **7,** (2012).

27. Kolaczkowski, B, Kern, AD, Holloway, AK & Genetics, B.-D. Genomic Differentiation Between Temperate and Tropical Australian Populations of Drosophila melanogaster. *Genetics* (2011). at <http://www.genetics.org/content/187/1/245.full-text.pdf+html>

28. Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N. & Quince, C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC bioinformatics* **17,** 125 (2016).

29. Huang, W, Massouras, A, Inoue, Y & Peiffer, J. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome …* (2014). at <http://genome.cshlp.org/content/24/7/1193.short>

30. Comeron, J. M., Ratnappan, R. & Bailin, S. The many landscapes of recombination in Drosophila melanogaster. *PLoS genetics* **8,** e1002905 (2012).

31. Rockman, MV & Kruglyak, L. Recombinational landscape and population genomics of

Caenorhabditis elegans. *PLoS genetics* (2009). doi:10.1371/journal.pgen.1000419
32. Zdraljevic, S, Roberts, JP & acids , A.-E. CeNDR, the Caenorhabditis elegans natural diversity resource. *Nucleic acids …* (2017). at <https://academic.oup.com/nar/article-abstract/45/D1/D650/2770657>
33. Nouhaud, P, Tobler, R & Nolte, V. Ancestral population reconstitution from isofemale lines as a tool for experimental evolution. *Ecology and …* (2016). doi:10.1002/ece3.2402
34. Lack, JB, Lange, JD & biology …, T.-A. A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus. *Molecular biology …* (2016). doi:10.1093/molbev/msw195
35. Teotónio, H, Chelo, IM, Bradić, M, Rose, MR & Long, AD. Experimental evolution reveals natural selection on standing genetic variation. *Nature genetics* (2009). at <https://www.nature.com/articles/ng.289>
36. Pettersson, ME & Honaker, CF. Standing genetic variation as a major contributor to adaptation in the Virginia chicken lines selection experiment. *Genome …* (2015). at <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0785-z>
37. Turchin, MC, Chiang, C. & Palmer, CD. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature …* (2012). at <https://www.nature.com/articles/ng.2368>
38. Kessner, D. & Novembre, J. forqs: forward-in-time simulation of recombination, quantitative traits and selection. *Bioinform Oxf Engl* **30,** 576–7 (2013).

# Methods

**Establishment and sequencing of founder set**

209 iso-female lines were established from populations sampled along a latitudinal cline in North America from Maine and PA. These isogenic lines were inbred for ~20 generations of full-sib mating to produce viable, fertile inbred lines. See Bergland et.al (in progress) for full details. 30-50 individuals from each line were pooled for DNA extraction. Whole flies were homogenized with lysis buffer and 1mm beads and DNA was precipitated from the homogenate before resuspension in TE buffer. Libraries were prepared with a modified Nextera protocol developed by (Baym, et al 2015). All samples were indexed with Illumina's TruSeq Dual Index Sequencing Primer Kit (PE-121-1003) and pooled equimolarly into 3 sets of ~70 samples each. Each set of pooled DNA libraries were purified using Ampure XP and size-selected to 450-500 bp with a SizeSelect E-Gel. After an additional 5 rounds of PCR, DNA libraries were purified using Ampure XP beads, quantified, and diluted to the appropriate concentration before sequencing on the HiSeq 3000. Adapter sequences were trimmed (Trimmomatic v0.33) and overlapping reads were merged (PEAR v0.9.6), then reads were mapped (bwa v 0.7.9) to the *D.melanogaster* reference genome (v5.39) using default parameters. PCR duplicates were removed using PicardTools (v1.12). Base quality score recalibration, indel realignment, and novel SNP discovery were carried out using GATK's Unified Genotyper (version 3.4-46), on all sequenced inbred lines together with a larger set of previously sequenced *Drosophila melanogaster* lines derived from multiple Europe and North America populations. Only SNPs segregating in the 99 inbred strains pooled for re-sequencing were used to simulate reads and estimate haplotype frequencies.

**Generating Experimentally Pooled Samples**

One male each was selected from each of 99 inbred strains, and all individuals were pooled for re-sequencing. A second biological replicates was constructed from 99 additional individuals. DNA isolation was performed as described above. 3 separate libraries were prepared from each of the two biological replicates using different library prep methods: [1] according to protocols described in Nextera DNA Library Prep Reference Guide (15027987 v01); [2] a modified Nextera protocol (as described above); [3] a Covaris shearing protocol. Final results from the 3 library prep methods were similar. All libraries were size-selected and PCR amplified using two replicate PCR reactions and a high volume of template DNA to prevent PCR-jackpotting. DNA was purified, quantified, and diluted before sequencing on the HiSeq 3000. Raw, 150bp pair-end reads were trimmed for adapter sequences with Skewer (version 0.1.127). Read merging, mapping, and PCR duplicate removal was performed as above.

### Generating Simulated Pooled Samples

150-bp paired end pre-aligned reads were simulated from a table of alternate founder genotypes and the *D. melanogaster* reference genome with simreads, a software tool included with the HARP package. All reads were simulated with an error rate of 0.2%. No read trimming or PCR duplicate removal was done. All SNP tables used to generate reads underwent imputation before read simulation.

### Haplotype Frequency Estimation

All haplotype likelihoods and frequencies were estimated with HARP. Haplotype frequencies were evaluated in 1000kb, 100kb or 10kb window width sizes, with 100kb, 10kb, or 1kb window step sizes, respectively. SNP tables used to assign haplotype frequencies were re-imputed separately from any SNP table used to simulate reads. For reference, inferring haplotype frequencies for 99 founder lines at 283k segregating sites on chromosome 2L in 1000kb windows took 8 minutes and required 450Mb RAM for samples sequenced at 5x empirical coverage and took 15 minutes and required 860Mb RAM for samples sequenced at 10x. Using 100kb windows took 9.5 minutes / 70Mb and 17.5 minutes / 132Mb for 5x and 10x samples, respectively.

### HAF Estimations

Allele frequencies calculated using haplotype inference (HAFs) were estimated as the sum of founder haplotypes containing the alternate allele, each weighted by their average inferred haplotype frequency in all haplotype inference windows overlapping the site. Founder haplotypes with missing calls were given a fractional alternate allele count equal to the mean of called founders with alternate alleles.

### Accuracy Estimations Using Effective Coverage

Effective coverage was used as a metric to assess accuracy of all HAFs and non-HAFs. Effective coverage was calculated by equating the total theoretical binomial variance (*BV*) of the true frequencies given an average coverage *C* [ where BV= $\sum \frac{p(1-p\ )}{C}$ ] to the sum of the squared error (*SSE*) of estimated allele frequencies. Solving for *C* with $C = \frac{\sum p(1-p)}{\mathrm{SSE}}$ yields the theoretical coverage at which binomial sampling of reads would be expected to contain the observed amount of error from estimated frequencies.

### HAFs with an alternate founder set

SNPs were derived from 205 strains initially isolated from Raleigh, NC that were independently sequenced as part of freeze 2 of the *Drosophila* Genetic Reference Panel (DGRP) [29] . Genotype data was downloaded directly from http://dgrp2.gnets.ncsu.edu and read simulation, haplotype inference and effective coverage calculations were carried out as above.

### Recombination

Forward-in-time simulations of recombination were performed with the software tool forqs[38] using a *D. melanogaster* recombination map[30]. Recombination breakpoints were simulated for up to 200 generations of the same evolutionary trajectory in 10 replicate populations with a constant population size of 1000 individuals. When selected sites were added, each replicate contained the same parameters for selection but included different selected sites. 198 recombination breakpoints were used to construct ‘sampled’ sets of chromosome pairs from corresponding segments of the 99 individually sequenced founder haplotypes. Read simulations were performed as above with recombined sets of SNPs. True allele frequencies were calculated as above from the same set of SNPs that were used to generate simulated reads.

### Founder Ambiguity

To simulate reads with founder ambiguity, SNP calls from additional North American lines were added to calls from the 99 seasonal inbred lines pooled for sequencing. Additional sets of lines were then added to obtain SNP calls from 124, 149 and 209 lines respectively. SNP tables were also made by removing 25 and 50 lines from the original set of 99 lines. Each SNP table contained the same number of SNPs and was imputed as above. Reads were simulated from each new SNP table and further downstream analysis was conducted as described above.

# Supplemental Text

## *Incorporating uneven pooling of individuals produces more realistic estimates of true allele frequencies*

Our ability to measure the accuracy of HAFs and non-HAFs depends on our ability to determine the true contribution of each pooled individual. Since uneven pooling is a source of error known to affect pool-seq samples[12], we estimated the relative contribution of DNA from each individual by calculating the average genome-wide allele frequency at sites private to each founder. While each founder could be detected in the pool, we found substantial variation in their relative representation (Supp. Fig 2). 'True' frequencies for the experimental pooled sample were thus calculated by weighting founders known to contain the alternate allele by their estimated representation in the pool. We assessed whether these 'true' allele frequencies were better recapitulated by experimental reads than 'true' allele frequencies calculated without incorporating uneven pooling at all fully genotyped sites (both private and common). We found that the effective coverage using unevenly pooled weighted values (126x) was higher than the effective coverage assuming evenly pooled individuals (120x). We used these same estimates of uneven pooling to simulate reads in uneven proportions from different haplotypes for the synthetic sample as well.

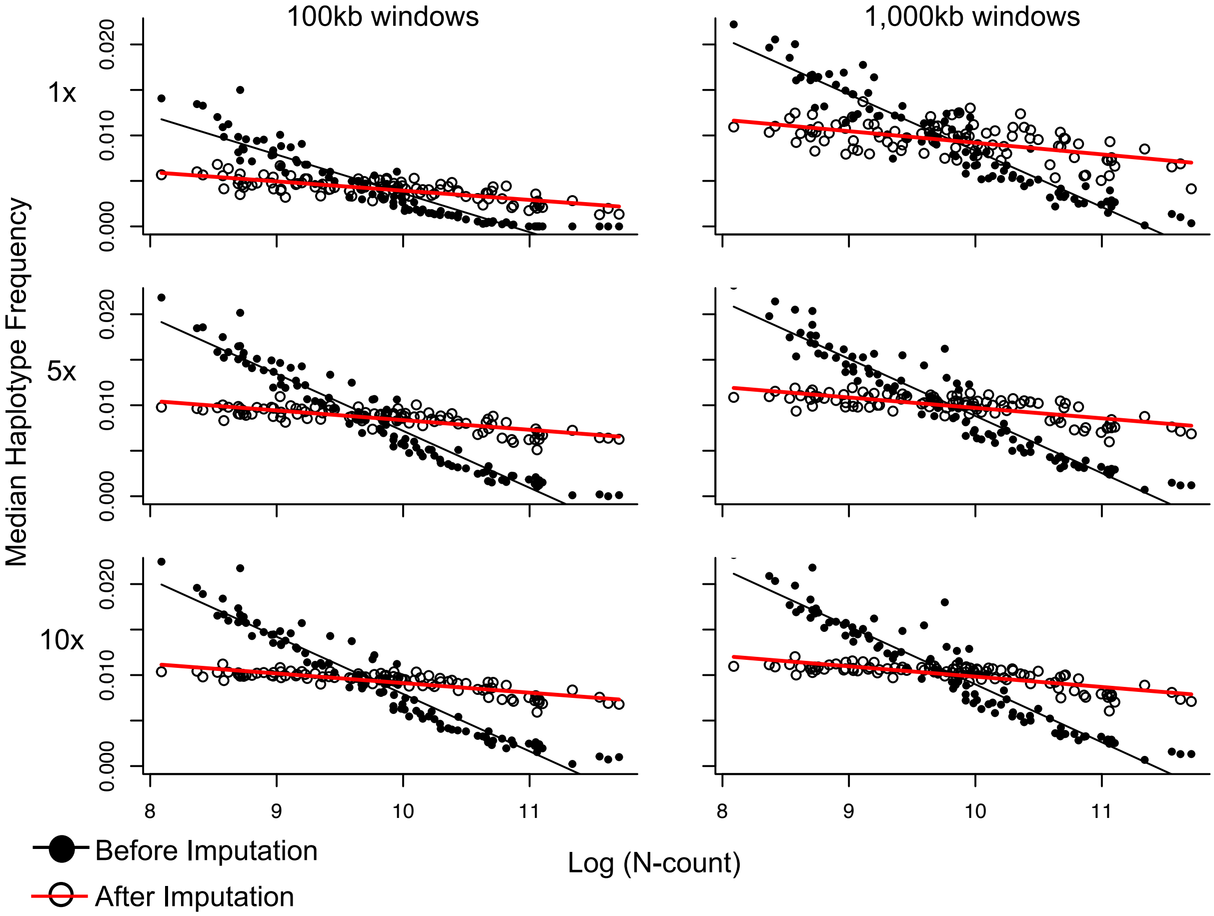
## *Imputing missing founder genotypes increases the accuracy of HAFs*

While missing information can be accommodated by many haplotype inference tools (i.e. an N in place of a missing call), it is unclear how missing calls affect inference accuracy, and what the best practices should be when missing calls are present in the reference founder set.

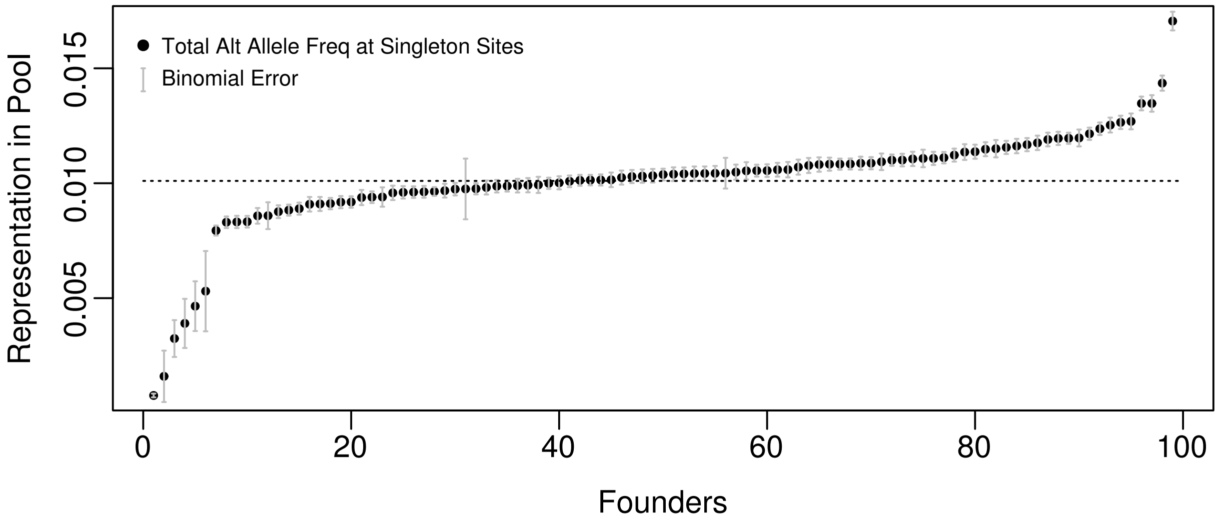
We first examined whether haplotype frequencies estimated for founders with many missing calls or few missing calls systematically deviated from an expected haplotype frequency of 0.101 (1/99). We found that across individual inference windows, there was a clear negative correlation between the number of missing calls per founder, and the haplotype frequencies estimated for that founder (Supp. Fig 1). To determine whether the observed skewed haplotype frequencies were directly associated with the presence of missing sites, we tested whether imputing genotype calls for missing sites would reduce bias in haplotype frequency assignment. To perform imputation, at each site we first calculated the allele frequency among called founder genotypes and used this value as a probability for assigning genotypes to missing calls. We found that imputation significantly reduced the skewed haplotype frequency distribution by 4-6-fold for all empirical coverages and window sizes tested.

We next examined how imputation of haplotype frequencies can impact the overall accuracy of HAFs. We found that imputation increased the accuracy of HAFs in all window sizes, but was most effective in 1000kb windows where accuracy increased by more than two-fold at higher empirical coverages (Supp. Table 2). We also confirmed that haplotype inference using imputed calls produced more accurate HAFs than using a subset of sites with no missing calls. Thus, we include imputation as a key step in our analysis pipeline.

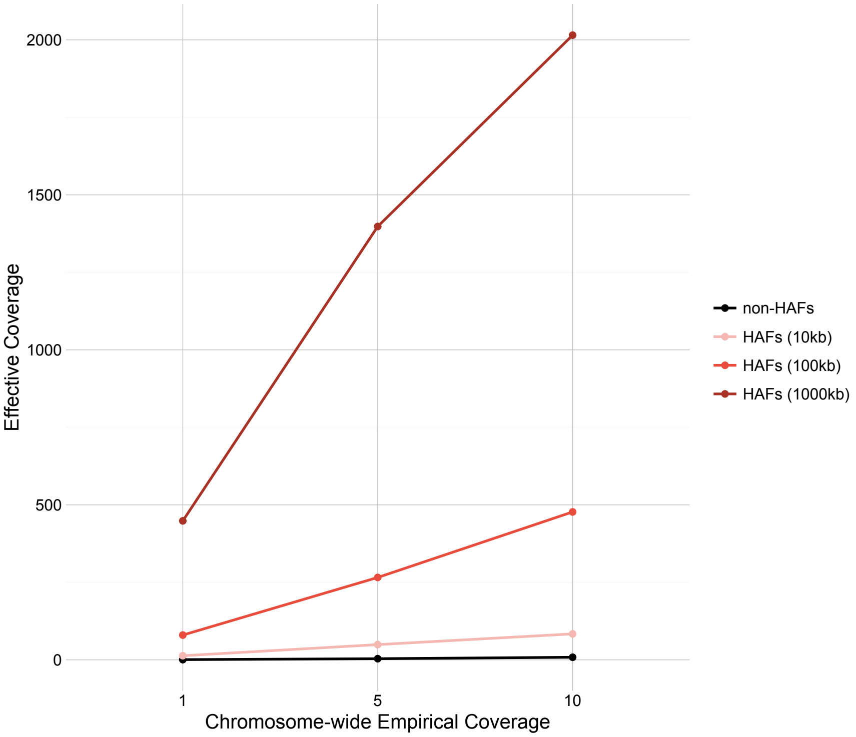
## Supplemental Figures and Tables



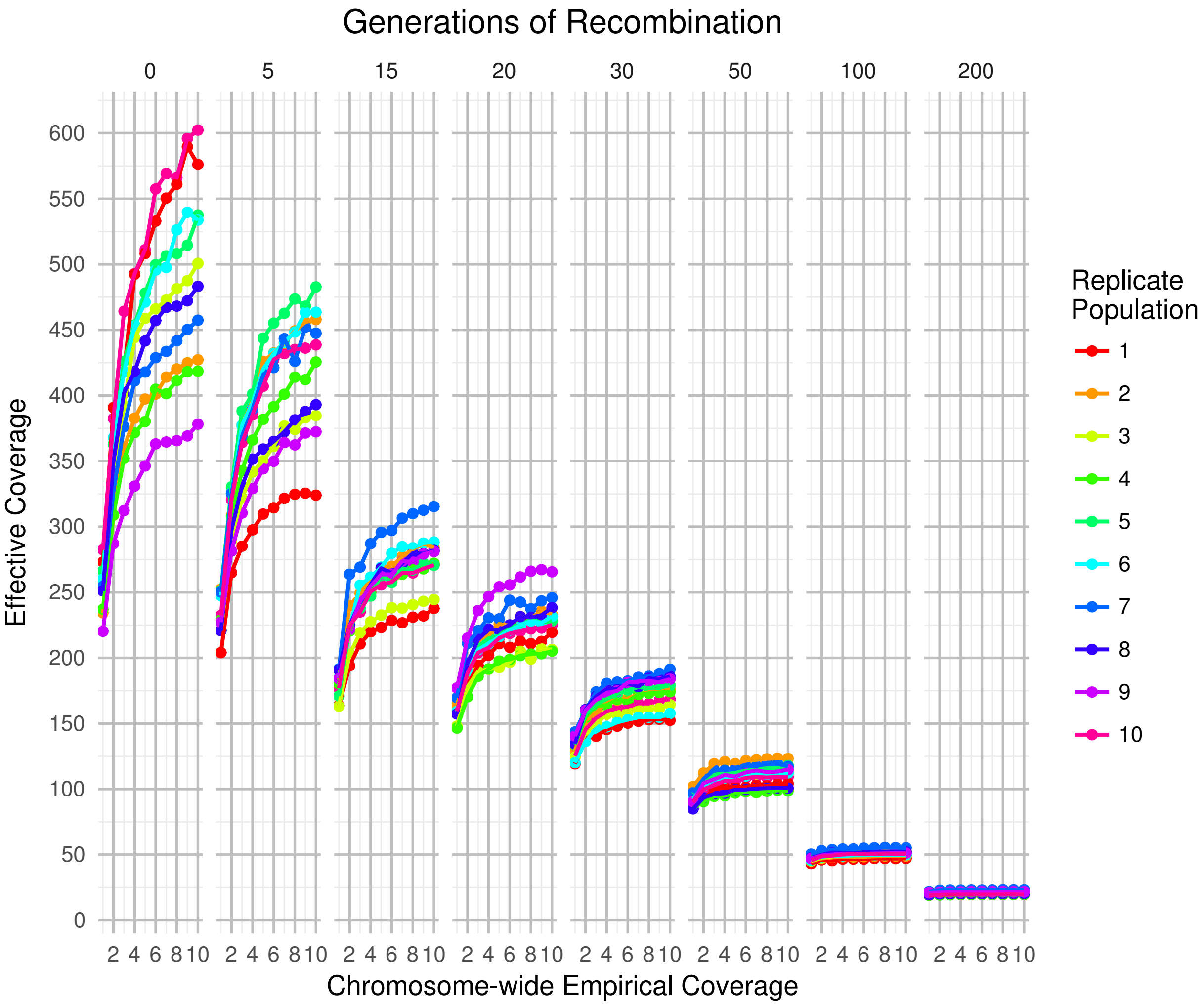
**Supplemental Fig 1**. Median haplotype frequency across all windows on chromosome 2L for each founder (n=99), calculated with different window sizes and empirical coverages. Haplotype frequencies calculated before imputation (filled circles) and after imputation (open circles) are plotted as a function of the log of the total number of ambiguous genotypes (aka "N-count"). Best fit lines for each dataset were calculated with standard linear regression.



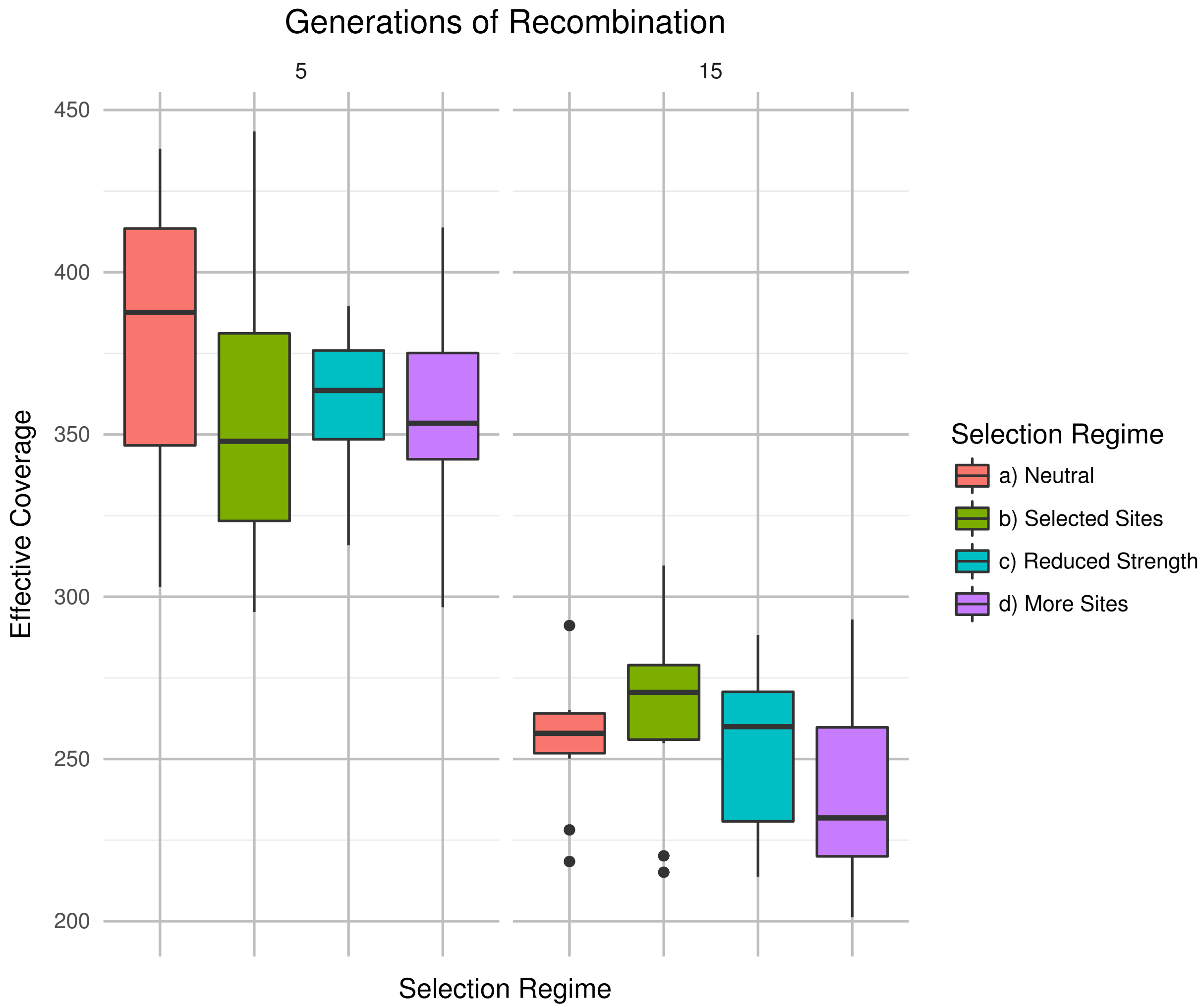
**Supplemental Figure 2.** Contribution of DNA from each pooled individual in experimental replicate 1, estimated by average genome-wide allele frequency across all singleton sites. The dashed line represents theoretical expectation for evenly pooled individuals. Points are colored by number of singletons sites per founder.



**Supplemental Figure 3.** Effective coverage values of HAFs using 1000kb (dark red), 100kb (red) and 10kb (pink) inference windows, and non-HAFs (black) within simulated pooled samples using a DGRP founder set (n=205) on chromosome 2L.



**Supplemental Figure 4.** Effective coverage values of HAFs calculated with 1000kb inference windows in 10 replicate populations after 0 to 200 generations of neutral recombination.



**Supplemental Figure 5.** Effective coverage of simulated samples from populations under a) the standard neutral selection regime, compared to populations under different selection regimes including b) 4 randomly distributed strongly selected sites contributing additively to a quantitative trait, c) same as panel b but with 50% reduced selection strength, and d) same as panel b but with twice as many selected sites. In each panel, effective coverage was calculated for 10 simulated replicate populations sampled at generations 5 (red boxplot) and 15 (yellow boxplot) and sequenced at an empirical coverage of 5x**.** Effective coverages from the same replicate are connected by a gray line.

| Window Size | Coverage | Ambiguous | Imputed | Subsetted |
|---|---|---|---|---|
| **10kb** | 1 | 9.7085 | 10.9811 | 10.5440 |
| | 5 | 34.0194 | 39.8801 | 37.3980 |
| | 10 | 58.4309 | 70.9542 | 63.7056 |
| **100kb** | 1 | 51.3650 | 60.5870 | 51.3393 |
| | 5 | 164.6956 | 215.1379 | 141.2237 |
| | 10 | 240.8699 | 345.4896 | 189.6102 |
| **1000kb** | 1 | 202.5682 | 329.7044 | 172.0310 |
| | 5 | 321.9890 | 701.5851 | 240.6798 |
| | 10 | 346.1369 | 809.1494 | 248.5619 |

**Supplemental Table 1**. Effective coverage values of HAFs obtained by different methods for treating missing founder genotype calls. Column 1 ('Ambiguous') refers to effective coverages obtained by performing haplotype inference with missing genotypes denoted by 'N'. Column 2 ('Imputed') refers to effective coverages obtained by performing haplotype inference with genotypes assigned to missing calls by randomly selecting the reference or alternate allele with a probability determined by the ratio of *called* ref and alt alleles at the site. Column 3 ('Subsetted') refers to effective coverages obtained by performing haplotype inference using only sites with full genotype information for every founder.