Title

Accurate, ultra-low coverage genome reconstruction and association studies in Hybrid Swarm mapping populations

Cory A. Weller[1,2], Alan O. Bergland[1]

[1] Department of Biology, University of Virginia, Charlottesville, Virginia 22904

[2] E-mail: caw5cv@virginia.edu

Abstract

Genetic association mapping studies seek to uncover the link between genotype and phenotype, and often utilize inbred reference panels as a replicable source of genetic variation. However, inbred reference panels can differ substantially from wild populations in their genotypic distribution, and patterns of linkage-disequilibrium and nucleotide diversity. As a result, associations discovered using inbred reference panels may not reflect the genetic basis of phenotypic variation in natural populations. To address this problem, we evaluated a mapping population design where dozens to hundreds of inbred lines are outbred for few (e.g. five) generations, which we call the Hybrid Swarm. The Hybrid Swarm approach has likely remained underutilized relative to pre-sequenced inbred lines due to the costs of genome-wide genotyping. To reduce sequencing costs and make the Hybrid Swarm approach feasible, we developed a computational pipeline that reconstructs accurate whole genomes from ultra-low-coverage (0.05X) sequence data in Hybrid Swarm populations derived from ancestors with phased haplotypes. We compared the power and precision of GWAS using the Hybrid Swarm, inbred lines, recombinant inbred lines, and highly outbred populations across a range of allele frequencies and effect sizes, modeling genetic variation from the Drosophila Genetic Reference Panel as well as variation from neutral simulations. While inbred populations tended to perform best due to the intrinsic power benefits conferred by the lack of heterozygotes, association mapping with the Hybrid Swarm performed comparably to highly outbred (F_{50}) populations and has higher precision than mapping with inbred lines. Taken together, our results demonstrate the feasibility of the Hybrid Swarm as a cost-effective method of fine-scale genetic mapping.

Introduction

Genetic mapping studies seek to describe the link between genotype and phenotype. For experimental crosses, mapping was traditionally conducted by scoring the phenotypes of recombinant offspring descended from a limited number of parental lines. While such QTL mapping studies can have high power to detect associations, they offer minimal mapping resolution (Cheng *et al.* 2010), often detecting broad regions of phenotypic association (Bergland *et al.* 2012). If linkage disequilibrium is lowered, spurious associations become rarer (Li *et al.* 2005) and associations can be resolved at the gene or nucleotide level, as in GWAS of large outbred populations (Nikpay *et al.* 2015; Wu *et al.* 2017; Monir and Zhu 2017). However, GWAS suffer from reduced power to detect associations, necessitating a large sample size relative to QTL mapping (Spencer *et al.* 2009).

To generate higher resolution mapping populations than the traditional biparental F2 design, Multiparent Populations (MPPs) are commonly used. By crossing together multiple inbred lines, researchers can produce genetically diverse mapping populations without sampling wild individuals. MPPs are commonly used for the dissection of complex traits in model organisms (Chesler *et al.* 2008; Kover *et al.* 2009; King *et al.* 2012b) and agriculturally important crops (Huang *et al.* 2012; Singh *et al.* 2013; Krämer *et al.* 2014). The mapping resolution of MPPs depends on the extent of linkage disequilibrium, and resolution is improved by allowing for more recombination between haplotypes, or by incorporating a greater number of genetically diverse haplotypes (Mott *et al.* 2000; Chia *et al.* 2005).

One alternative approach for generating a high-resolution mapping population is to substitute extensive recombination for increased haplotype diversity. By crossing dozens to hundreds of inbred lines for a limited number of generations, heterozygous mapping populations can be generated quickly with sufficiently reduced LD to detect associations with high resolution. Unfortunately, the many-haplotypes few-generations method is not without its drawbacks. First, including many haplotypes decreases the

frequency of the rarest alleles, reducing power to detect associations. Second, such an outbred population would require recurring genotyping efforts (Yang *et al.* 2018) unlike pre-sequenced homozygous lines. The net requirement of genotyping a large sample size may explain the widespread use of pre-genotyped inbred reference panels for genetic association experiments in model systems (Huang *et al.* 2011; King *et al.* 2012b; MacKay *et al.* 2012; Srivastava *et al.* 2017).

Here, we describe computational methods that allow for cost-effective association mapping with a large outbred population. The Hybrid Swarm is founded by dozens to hundreds of inbred lines, crossed for a limited number of generations. To reduce genotyping costs of the Hybrid Swarm, we developed and evaluated a pipeline to reconstruct whole genomes using ultra-low coverage sequencing data. We developed and tested our pipeline by reconstructing whole genomes for thousands of simulated Hybrid Swarm individuals. Our simulated genomes draw from natural variation in the *Drosophila melanogaster* Genetic Reference Panel (DGRP), as well as from variation generated from coalescent models representing a broad range of genetic diversity parameters for common model systems. We show that the Hybrid Swarm approach allows for highly accurate genotyping (average 99.9% genotypic accuracy) from ultra-low-coverage (0.005-0.05X) whole-genome individual-based sequencing. We then perform simulated GWAS to describe the power and precision of association mapping in the Hybrid Swarm compared to inbred lines, recombinant inbred lines, and a highly outbred (F_{50}) population. Our computational tools are capable of efficiently simulating low-coverage reconstruction and GWAS power analysis of any model system. Together, our results the feasibility of cost-effective high-resolution association mapping in a large outbred population.

Methods

Generating and preparing simulated reference panels.

In order to evaluate low-coverage reconstruction for various degrees of genetic diversity, we generated reference panels using haplotypes produced by coalescent

models across a range of genetic diversity levels. Haplotypes were generated using the *R* (R Core Team 2016) package *scrm* (Paul R. Staab et al. 2015) and subsequently restructured into VCF file format (Danecek et al. 2011). We generated ten independent panels for each of all 18 combinations of population size ($N_e = 10^4, 10^5, 10^6$), mutation rate ($\mu = 10^{-9}, 5 \times 10^{-9}, 10^{-8}$), and number of haplotypes (32, 128). The value for θ for each simulation was defined as $4N_e\mu$. We simulated a chromosome-length locus of 25 Mb with a recombination rate of 1.5 cM/Mb. SNP positions output by *scrm* (a decimal within the range of 0 to 1) were converted to base pair positions by multiplying the decimal by chromosome length (25 ×10^6 base pairs for our simulations) and rounding down to the nearest integer. Any sites with more than two alleles were converted to a biallelic site by discarding tertiary or quaternary alleles. Genotype values were re-coded as polarized signed integers: +1 for reference and -1 for alternate alleles. For every position, reference and alternate alleles were defined by randomly selecting one of the twelve non-repeating pairs of nucleotides. Reference genome FASTA files were created with a custom python script that generated a 25 million length string of nucleotide characters with weighted probability to achieve 45% GC-content, followed by replacing variable positions with their respective reference alleles.

Preparing DGRP haplotype data

As a case study of low-coverage genome reconstruction in a model system, we incorporated wild fruit fly genetic diversity from the Drosophila Genetic Reference Panel (MacKay *et al.* 2012) DGRP freeze 2 as available from the *Drosophila* Genome Nexus (Lack *et al.* 2015). To minimize missing data, we included the 129 lines (out of 205) which exhibited aligned whole genome FASTA files with less than 50% of nucleotides indicated by the ambiguity character N. We excluded insertions, deletions, fixed sites, and sites with more than two alleles. Any heterozygous genotype calls were masked as missing data. Diploid genotypes were re-coded as a single signed integer value, with +1 for homozygous reference, -1 for homozygous alternate, and 0 for missing data. This resulted in a polarized VCF file containing only biallelic SNPs and only homozygous (or missing) genotype calls.

Simulating Mapping Populations

To generate simulated populations, we developed a forward-simulator in R that stores ancestral haplotype block maps instead of genotypes. Our analyses necessitated a method of storing genotype information for thousands of individuals across thousands of simulations. To do so, we leveraged information redundancy that exists between related individuals in recombinant populations, generating haplotype block files. We achieved between three and four orders of magnitude of compression relative to a VCF file. For example, for a population containing 5000 diploid genotypes at nearly four million sites, a compressed VCF file is approximately 6.5 GB, compared to approximately 3.5 MB for a haplotype block file. This reduced file size is what allowed us to generate and store 28,000 total independent GWAS simulations (500 each for 56 parameter combinations). When haplotype block ancestry is known and recorded, as is possible with simulations, genotypes must only be recorded once (for the ancestral founders). Recombinant individual genotypes can then be reconstituted by extracting ancestral genotypes from ancestor and base pair position indices.

We simulated Hybrid Swarms through random mating over five non-overlapping generations at a population size of 10,000. Simulations proceeded in the following manner: first, a subset of either 32 or 128 founders was selected. Then, of that founder subset, 10,000 individuals were sampled with replacement. All possible founders were chosen with equal probability and assigned male or female sex with a 1:1 ratio, where sex was determined by the presence of a designated sex chromosome. Sexual reproduction was simulated by random sampling of recombinant gametes from male-female pairs. Once 10,000 recombinant progeny were generated, the parental generation was discarded. Reproduction continued until the F_5 population was achieved. Recombination frequency was modeled as a Poisson process with an expected value $\lambda = \Sigma(Morgans)$ per chromosome. For simulations of Drosophila populations based on DGRP chromosomes, recombination occurred only in females, with recombination frequency and position based on values from Comeron et al (2012). For populations founded by simulated haplotypes, recombination occurred in both

sexes, with recombination occurring uniformly across each chromosome (Supplemental Figure S1).

Simulating and Mapping Sequencing Data

We used *wgsim* (Li 2011) to generate simulated reads. To achieve a desired level of sequencing coverage $C = 0.05$ or 0.005, we generated $N = (C \times S)/(2 \times L)$ reads per chromosome, with read length $L = 100$ bp and chromosome length S bp. We specified a base error rate of 0.001 and an indel fraction of 0. Remaining *wgsim* parameters were left as default.

We assembled paired end reads using *PEAR* (Zhang *et al.* 2014) and separately aligned the assembled and unassembled groups to a reference genome with *bwa* 0.7.14 using the BWA-MEM algorithm (Li 2013). Reads from DGRP-derived populations were mapped to the *D. melanogaster* reference genome v5.39, and reads from coalescent-derived populations were mapped to their respective simulated reference genomes. After converting mapped reads to compressed BAM format with *samtools* 1.3.1 (Li *et al.* 2009), we removed PCR duplicates with *Picard tools* 2.0.1 (Broad Institute 2015a).

Most Likely Ancestors Selection

To make chromosome reconstructions in the hybrid swarm computationally tractable (Figure 1), we developed a method of accurately selecting a subset of most likely ancestors for any single chromosome. We then used that ancestor subset to reconstruct haplotype blocks using the *RABBIT* package (Zheng *et al.* 2015) in Mathematica. RABBIT operates as a Hidden Markov Model (HMM) using the Viterbi algorithm to return the most likely series of parental combinations (hidden states) across the genome (SNP positions) given the observations (sequenced alleles). For every position in the genome, the Viterbi algorithm evaluates relative likelihoods of transitioning to any possible hidden state. Because the hidden states in our case are ancestor combinations, there will be $(N^2 + N)/2$ combinations of N haplotypes to evaluate at every site. This number of evaluations is tractable at smaller values of N but grows at a

quadratic rate. For example, increasing the number of founding haplotypes from 8 to 128 is a 16-fold increase in haplotypes, but it would incur orders of magnitude increases in computational effort (Figure 1). Thus, in order to make reconstructions in *RABBIT* computationally tractable for hybrid swarm individuals, it is necessary to identify a subset of founders that accurately includes the true ancestors contributing to any given chromosome.

We used the software package *HARP* (Kessner *et al.* 2013) to rank the population founding lines based on likelihood of being a true ancestor of a chromosome to be reconstructed. *HARP* was originally developed to estimate haplotype frequencies from pooled sequence data, and we co-opted it to assess relative likelihood that any founder contributed to a genomic window. We ran *HARP* with non-overlapping 100 kb windows with a minimum frequency cutoff 0.0001, producing output which can be visualized as a heat map of ancestor likelihood across the chromosome. A custom *R* script analyzed this *HARP* output and ranked all possible founders in terms of likelihood of contribution for a given chromosome. Briefly, a chromosome-wide significance threshold was calculated, e.g. the 95% or 99% quantile of all likelihoods across all founders and all chromosome windows. Then, every potential ancestor for each 100 kb window was classified as falling above or below this threshold. Founding lines were then ranked in descending order of the number of windows passing the threshold. We examined two measures of effectiveness for this method across a range of quantile threshold values (90%, 95%, 99%, and 99.9%) when selecting up to a maximum number of most likely ancestral haplotypes. The first measure is the number of true ancestral founders excluded; the second measure is the fraction of the chromosome derived from ancestors missing from the selected subset.

Chromosome Reconstruction with RABBIT

We used the MAGIC reconstruct method of the Mathematica package RABBIT (Zheng *et al.* 2015) to perform chromosome reconstructions, which has been shown to be accurate for genotype estimation at sequencing coverage at 0.05X for a variety of multiparent populations (Zheng *et al.* 2018). RABBIT requires three inputs: observed

genotypes in the individual being reconstructed; map distance (in cM units) of the same loci; and genotypes for the potential ancestors at those same loci. For DGRP-derived simulated populations, we specified map distance based on values reported by Comeron et al. (2012) by performing linear interpolation of cumulative map units (cM) as a function of base pair position. For populations derived from simulated haplotyopes, we used a linear function of 37.5 cM over each 25 Mb chromosome. To specify genotype information, we first counted reference and alternate reads using the Genome Analysis Toolkit *ASEReadCounter* (Broad Institute 2015b). Because it is not possible to make confident homozygote genotype calls from low coverage sequencing data where most sites are observed only once and or twice, we only included diploid genotype observations for sites where both reference and alternate alleles were observed. As RABBIT allows for an ambiguous allele character, for all sites where only reference or alternate reads were observed (but not both), we included one ambiguous allele.

To minimize memory and runtime requirements, we included at most 5,000 SNPs per chromosome, selected for maximum ancestor-discerning information content. If an observed (sequenced) allele is common, it will only slightly narrow down the possibility of ancestors. If a sequenced allele is rare—at the most extreme, unique to one individual—it provides greater information from which founder that site is derived. Thus, we designate information-rich sites as those where the frequency of the sequenced allele is the lowest with respect to the pool of most likely ancestors. In order to sample sites with high information content spread throughout the chromosome, we used an iterative approach. First, we included all heterozygous sites (i.e. where reference and alternate alleles are both observed). Then, 10% of all SNPs were randomly sampled, and we retained up to the top 0.2% most informative sites, repeating the random sampling and retention until we designated 5,000 SNPs.

We ran RABBIT independently for each chromosome using the Viterbi decoding function under the joint model, with all other RABBIT parameters left at default. RABBIT output was converted to a phased chromosome haplotype map, which we then used to extract and concatenate genotype information from a VCF file containing founder

genotypes. To calculate genotype reconstruction accuracy, we first imported true (simulated) and estimated (reconstructed) genotypes using a custom *R* script. We measured the fraction of all remaining sites where the estimated diploid genotype is identical to the originally simulated diploid genotype, excluding fixed sites with respect to the founding haplotypes, and excluding any sites with missing genotype information. Because male individuals do not possess two copies of the sex chromosome, we only evaluated accuracy for autosomes.

To measure accuracy of estimated frequency of recombination events, true and estimated recombination counts were first summed over both copies of each chromosome in a simulated individual. This removed the possibility of introducing error by comparing the wrong copies of chromosomes. Only detectable recombination events were considered, i.e. those that did not occur between homologous haplotypes. We then used the *epi.ccc* function of the *R* package *epiR* (Stevenson 2018) to calculate Lin's concordance correlation coefficient (ρ) between the true and estimated recombination counts.

Modeling Computational Complexity of Chromosome Reconstruction

To estimate the rate at which computational requirements grows with data input, we performed chromosome reconstructions with varying numbers of potential founders and markers (SNPs). This allows us to extrapolate the runtime and memory for performing the most resource intensive chromosome reconstructions (i.e. those with > 40 founding lines). To generate runtime and memory usage data, we performed 900 reconstructions using varying sizes of RABBIT input for a single example individual 2L chromosome. Reconstructions included the four true ancestors of the simulated individual, plus 0 to 32 additional haplotypes (for a total of between 4 and 36 founders, in steps of 4) and a random selection of marker SNPs (between 500 and 5000 in steps of 500). Ten replicates, each with a unique random set of SNPs, was conducted for each combination of N founding lines and S SNPs using a single core on the University of Virginia computing cluster, with total runtime and peak memory usage as reported from the SLURM workload manager (CPUtime and MaxRSS, respectively). We then

modeled the mean runtime and memory usage (averaged across 10 replicates per parameter combination) as a function of number of founding lines and number of SNPs fed into RABBIT. For runtime, simulations involving 8 or fewer founding lines were omitted from the regression model because they ran too quickly to resolve non-zero runtime. Memory was modeled as $Memory(GB) = 7.367 \times 10^{-9} \times SN^4 + 0.0316$, while runtime was modeled as $Runtime(Minutes) = [1.189 \times 10^{-3} \times N^2 + 1.038 \times 10^{-6} \times SN^2 + 2.649 \times 10^{-4} \times S]^2$.

Simulated GWAS

We performed GWAS on mapping populations produced by random sampling and permutation of the previously-described forward-simulated populations. Although the forward simulator we developed is efficient, it would not have been computationally feasible to simulate 500 fully independent mapping populations (per parameter combination) in a reasonable amount of time. Instead, we generated ten independent forward-simulated populations, and for each of those, generated fifty randomly permuted subsets (Figure 2). For a single simulated mapping population, we began by sampling (with replacement) a random subset of 5,000 individuals, out of 10,000 total individuals generated by forward-simulation. Then, we performed a permutation of haplotype ancestry with a new, randomly-ordered (equally sized) subset of founders. The permutation of ancestry was one-to-one, e.g. all haplotype blocks that were previously derived from founder X would be translated to founder Y, and blocks previously derived from Y would in turn be mapped to founder Z.

In addition to Hybrid Swarm populations, which we ran through the simulated sequencing and mapping pipeline, we generated four additional types of mapping populations for comparing GWAS performance: Highly outbred (F_{50}) populations, similar to sampling wild individuals; Inbred Lines (ILs), similar to mapping with the DGRP); and Recombinant Inbred Lines (RILs), similar to mapping with the DSPR.

The F_{50} populations were generated in same manner as the Hybrid Swarm, except for fifty non-overlapping generations of recombination instead of five generations. The ten

resulting forward-simulated populations were resampled and permuted as we did with the Hybrid Swarms.

We simulated ten initial sets of 800 RILs using the same forward-simulator as previously described, each initialized with a random subset of eight DGRP haplotypes. Populations randomly recombined at a population size of 10,000 for fifty non-overlapping generations, after which 800 random male-female pairs of individuals were isogenzied through 25 generations of full-sibling mating. This scenario roughly corresponds to the Drosophila Synthetic Population Resource (King *et al.* 2012a). For computational simplicity, after the 25 generations of isogenization we removed any remaining residual heterozygosity by forcing the identity of a second chromosome copy to be identical to the first copy. We then sampled 5,000 draws (with replacement) of the 800 RILs followed by ancestry permutation as described above.

To simulate GWAS on Inbred Lines, no forward-simulation was necessary. For a single simulated population, we first randomly selected 128 DGRP lines with high coverage and low levels of heterozygosity as the set of founders. Then, those 128 lines were randomly sampled with replacement 5,000 times. As with hybrid swarm and RILs, for any parameter combination we generated a total of 500 mapping populations.

Phenotypes were modeled as probabilistic assignment to a case or control group dependent on allele dosage at a purely additive single SNP. We designated a causal locus as a random autosomal biallelic SNP segregating within 0.5% of a desired minor allele frequency (50%, 25%, and 12.5%). We modeled SNPs at 5% and 10% percent variation explained (PVE), where reference allele homozygotes were assigned to the case group with probability $50\% - PVE/2$, and alternate allele homozygotes were assigned to the case group with probability $50\% + PVE/2$. Heterozygotes were equally likely to be assigned case and control.

To perform many replicates of GWAS for many parameter combinations, we performed a simple χ^2 test of independence for reference and alternate allele counts between case

and control groups. To do so most efficiently, we developed a method of aggregating allele counts that uses a haplotype map table in conjunction with a single table of founder genotypes (Figure 2). Briefly, haplotype table breakpoints across all individuals were sorted in ascending order. When iterating through ascending unique start and stop positions, between any pair of breakpoints, all SNPs will be comprised of the same number of each founding haplotype. Haplotype IDs could then be counted and sorted in the same column position order as the table containing polarized allele status (-1 for alternate, +1 for reference). Multiplying the genotype table by the haplotype count vector results in final allele counts, polarized negative for alternate alleles and positive for reference alleles. For inbred mapping populations, we corrected for non-independent allele draws by dividing the χ^2 value by two.

To describe the accuracy of simulated GWAS, we measured the likelihood of including a locus that is near the causal site when considering a set of the top N most significant SNPs. Here, 'near' is defined as either exact-SNP resolution, or within 1, 10, or 100 Kb. In the case of 1 kb precision, we first consider the set of SNPs +/- 1 kb from the most significant locus (greatest chi-square statistic). Then, we consider the second set of SNPs as those within +/- 1 kb of the most significant locus outside of the window already accounted for. This selection of significant clusters was repeated iteratively for the top 25 regions, for window sizes of 0 (exact SNP resolution), 1 kb, 10 kb, and 100 kb.

We calculated genomic inflation factor (GIF, λ_{1000}) as the value of $\chi^2_{observed}/\chi^2_{expected}$ with two degrees of freedom. Because GIF increases with sample size, we performed a correction to report the level of GIF expected with a sample size of 1,000 case and 1,000 control individuals (Freedman *et al.* 2004).

Assessing counts of variable sites at appreciable frequency in the DGRP

There is a reduction in power to detect associations with alleles segregating at low minor allele frequencies. When a population is founded by N lines, any SNP will be segregating at a relative frequency of at least 1/N, given that the SNP is not fixed within

the population and haplotypes are equally represented. We counted the number of sites on a given chromosome arm segregating above a minor allele frequency threshold of MAF=(0.05, 0.125, and 0.25) for random draws (without replacement) when sampling N=(2, 4, 8, 16, 32, 64, 128) haplotypes of the 129 included DGRP lines We performed this sampling 20 times for each chromosome arm.

Data Availability Statement

The code used to generate, process, and plot our data is available on GitHub: https://github.com/cory-weller/low-coverage-genome-reconstruction

Results

Computational Complexity of Chromosome Reconstruction

To determine reasonable limits for numbers of SNPs and haplotypes used for chromosome reconstruction with RABBIT, we modeled peak memory usage and runtime across a range of input sizes. Peak memory grew linearly with number of SNPs used, and at a greater-than-linear rate with haplotypes (Figure 1A, $Memory = 7.367 \times 10^{-9} \times SN^4 + 0.0316)$, $F = 3.534 \times 10^6$, $df = 1 \& 88$, $R^2 = 1$). The runtime of RABBIT increased at a greater-than-linear rate for both number of SNPs and number of haplotypes, though the *N* parameter dominates (Figure 1B, $Runtime = [1.189 \times 10^{-3} \times N^2 + 1.038 \times 10^{-6} \times SN^2 + 2.649 \times 10^{-4} \times S]$, $F = 4.316 \times 10^4$, $df = 3 \& 67$, $R^2 = 0.9995$). These models allowed us to estimate resource requirements at greater numbers of haplotypes (Figure 1, C & D) which would be unfeasible to measure empirically.

Most-Likely-Ancestor Selection

To reduce computational requirements of haplotype reconstructions with RABBIT, we developed and evaluated an algorithm for selecting a minimum representative set of Most-Likely-Ancestors (MLAs) for chromosome reconstruction. We found a HARP threshold of 0.99 (see methods) discerned a minimal subset of founding lines that tended to include a given chromosome's true ancestors (Figure 3). At this threshold,

outcomes became asymptotic at the computationally tractable cap of 16 founding lines (Figure 1). Thus, we performed chromosome reconstruction using up to 16 most-likely-ancestors as inferred with a HARP threshold of 0.99.

In all cases, decreasing the HARP threshold from 0.95 to 0.90 further reduced chromosome representation while increasing the number of extraneous founding lines selected for reconstruction. While a higher HARP threshold of 0.999 yielded the smallest and most computationally tractable set sizes of MLAs ($\overline{\mathrm{N}}$=2.4-3.5), the strict threshold excluded true ancestors, resulting in a set that is least representative of chromosomes to be reconstructed. For 128-founder populations, a threshold of 0.999 failed to identify founders constituting an average of 3.33% and 13.9% of chromosomes for DGRP- and Coalescent-founded populations, respectively. In 32-founder populations, the 0.999 threshold missed founders representing an average of 15.5% and 29.7% of chromosomes from DGRP- and Coalescent-founded populations, respectively.

Populations simulated with genetic variation derived from coalescent models described above included the parameters $N_e = 10^6$ and $\mu = 5 \times 10^{-9}$. The effectiveness of most-likely-ancestor selection for populations modeled across extended values of N_e and μ is shown in supplemental Figures S2 and S3, respectively. Similarly, the number of most-likely-ancestors chosen for reconstruction in RABBIT are shown in Figures S5 and S6.

Selected MLA set size is described in Figure S5 for 32-founder populations and Figure S6 for 128-founder populations. Ancestor selection effectiveness for DGRP-derived populations at two levels of sequencing coverage (0.005X and 0.05X) is shown in supplemental Figure S4, and the corresponding number of most-likely-ancestors chosen for reconstruction are shown in supplemental Figure S7.

Reconstruction Accuracy

Chromosome reconstruction of simulated F_5 Hybrid Swarm genomes at 0.05X sequencing coverage yielded highly accurate genotype estimates (Figure 4). The

median percent of sites with correctly estimated genotypes was greater than 99.9% whether the population was founded by 32 or 128 founding lines for either DGRP or coalescent ($N_e = 10^6$ and $\mu = 5 \times 10^{-9}$) haplotypes. We additionally report reconstruction accuracy in coalescent-derived populations across a range of N_e and μ values in supplemental Figure S8.

For simulations founded by DGRP lines, 80.5% of reconstructed chromosomes from 32-founder populations exhibited > 99.9% accuracy, with the remaining 19.5% of reconstructions contributing to a long tail with a minimum of 84.37%. Increasing the number of founding lines to 128 resulted in genotype accuracy above 99% for all cases (minimum: 99.4%), with 83% of reconstructed chromosomes achieving greater than 99.9% accuracy.

Although median accuracy for coalescent-derived populations was equivalent to that of DGRP-derived populations (99.9%), coalescent-derived populations with 32 founders exhibited a greater number of low-accuracy reconstructions. While 82.5% of simulations with 32 coalescent haplotypes were at least 99% accurate, the remaining 17.5% of reconstructions contributing to a long tail with a minimum accuracy of 59.7%. Increasing the number of founding lines to 128 resulted in 96.3% of simulations being greater than 99% accurate, with a minimum accuracy of 89.6%.

The number of recombination events estimated from chromosome reconstruction was most accurate for populations founded by 128 lines (Table 1). Reconstructions of DGRP- and Coalescent-derived chromosomes yielded recombination count estimates that were 98.6% and 95.6% concordant with their respective true recombination counts (Lin's concordance correlation coefficient, ρ). When populations were founded by 32 lines, recombination count estimates were more inaccurate, with DGRP- and Coalescent-derived reconstructions achieving 50.2% and 75.9% concordance with their respective true recombination counts. For 32-founder populations, DGRP-derived reconstructions tended to slightly overestimate recombination counts, while the same counts were underestimated for coalescent-derived populations.

Simulations that inferred an unlikely high number of recombination events tended to exhibit reduced accuracy (Figure 4). All DGRP-derived simulated individuals (of 1600 total) exhibited ≤ 8 recombination events, and all but three coalescent-derived simulated individuals (7197 of 7200 total) exhibited ≤ 9 recombination events. Accordingly, we considered any reconstructions to be 'hyper-recombinant estimates' if the inferred recombination count is greater than 8 for DGRP-derived populations or greater than 9 for coalescent-derived populations.

At 0.05X sequencing coverage, hyper-recombinant estimates did not occur for 128-founder populations, and only rarely resulted from 32-founder populations. Within DGRP-derived 32-founder populations, reconstructions with hyper-recombinant estimates were below the sixth percentile of genotype accuracy (N=6/400 simulations, genotype accuracy range=92.8%-98.8%). For coalescent-derived 32-founder populations, reconstructions estimated as hyper-recombinant fell in the bottom 9% of genotype accuracy (N=3/400 simulations, genotype accuracy range = 92.1%-95.6%). Although hyper-recombinant estimates always fell in the bottom 10% of accuracy, the least accurate reconstructions were not hyper-recombinant. For coalescent-derived 32-founder populations, 4.25% (17/400) of reconstructions without hyper-recombinant estimates exhibited lower genotype accuracy than the least accurate hyper-recombinant simulation (range = 59.7%-92.1%). Similarly, for DGRP-derived 32-founder populations, 2.5% (10/400) of reconstructions without hyper-recombinant estimates exhibited lower genotype accuracy than the least accurate hyper-recombinant simulation (range = 82.3%-92.8%).

Reducing sequencing coverage by an order of magnitude from 0.05X to 0.005X resulted in more frequent hyper-recombinant reconstruction estimates, though overall median genotype accuracy remained above 99% (Figure S9). Hyper-recombinant reconstructed chromosomes exhibited genotype estimates with accuracy below 99%, while the remaining simulations (with lower recombinant counts) achieved above 99% genotype accuracy. For populations founded by 32 DGRP lines and sequenced at 0.005X

coverage, 14% of simulations produced hyper-recombinant estimates (N=56/400), of which only 26.8% (N=15/56) surpassed 99% genotype accuracy (median=98.5%). The remaining 86% of simulations (N=344/400) that were not hyper-recombinant retained greater accuracy, with 89% of simulations resulting in at least 99% genotype accuracy (median=99.5). Increasing the number of founding DGRP lines from 32 to 128 at 0.005X coverage failed to eliminate hyper-recombinant estimates. With 128 founding lines, 14.5% of simulations were hyper-recombinant (N=58/400), of which 24.1% (N=14/58) surpassed 99% genotype accuracy (median=98.6%). The 85.5% of simulations that were not hyper-recombinant (N=342/400), exhibited accurate genotype estimates, with 86.5% (296/342) of simulations achieving over 99% genotype accuracy (median=99.6).

GWAS Simulation Accuracy

To report the power of a GWAS, we must first define a “true positive” result. Consider a putative SNP identified by GWAS that is 50 kb from the causal SNP. Such a result would be considered a false positive if the aim is to identify the exact responsible nucleotide, but may be a true positive with respect to to identify an associated gene. To cover both use cases, we describe a true positive in terms of both SNP-level resolution (requiring an exact base-pair match), or region-level resolution (allowing for tolerance up to 100kb between putative hits and the causal SNP). Additionally, it is unrealistic to simply evaluate the single top result of a GWAS. Rather, a set of candidate loci may be chosen for follow-up evaluation in confirmatory studies, and the probability of including the causal SNP will increase as a greater number of putative SNPs are evaluated. Most distinct changes in GWAS power occurred when including between most significant to top 10 most significant candidate loci, after which power increased at a reduced rate, if an asymptote was not already reached.

The estimated power of GWAS using a specific type of mapping population, i.e. the fraction of simulations with a true positive, is shown in Figure 5. For simplicity, we focus on GWAS power when including the top 10 most significant candidate loci—a reasonable number of putative sites that may be investigated in follow-up studies.

$_5$ Hybrid Swarms founded by either 32 or 128 founding lines exhibited nearly equivalent power compared to $_{50}$ outbred populations across all parameter combinations. For common alleles, i.e. those segregating at 50% frequency, all outbred populations achieved approximately 50% power to identify a causal variant with SNP-level precision, and 70% power at the gene-level. Both inbred populations were highly effective at detecting associations at the gene-level (99% and 99.8% for ILs and RILs, respectively). SNP-level power was one fourth lower than gene-level power for RILs (75.4%), but only marginally reduced for inbred lines (97.4%).

Power to detect associations is reduced when the causal allele is rare (segregating at 12.5% frequency). For such rare alleles, the gene-resolving power of ILs drops by nearly half (to 54.8%), while RILs maintained high power (81.8%). All outbred populations exhibited approximately 20% power to detect rare alleles at the gene-level. Inbred lines were the sole frontrunner for identifying low frequency alleles with SNP-level resolution (37.2%), followed by 128-founder Hybrid Swarm (10.8%), F50 outbred (8.2%), 32-founder Hybrid Swarm (6%), and RILs (3.2%).

GWAS Genomic Inflation Factor

If individuals are assigned to case and control groups with equal probability, then the resulting χ^2 statistics should follow the expected distribution. If individuals are not sorted into groups randomly, i.e. allele state at a causal SNP dictates nonrandom group assignment, then χ^2 values for that SNP should be inflated to some extent. Nonrandom associations between a causal SNP and other loci can inflated test statistics across a chromosome, or across a whole genome. The genome-wide inflation factor (λ) can be expressed as the ratio of observed and expected median χ^2 values (Figure 6). Because our simulations model a single causal SNP, λ is a reflection of greater-than-chance associations arising due to linkage with the causal SNP being modeled, which can serve as a proxy for false positive rate.

Because the median expected χ^2 statistic increases with sample size, we report λ_{1000}, a sample-size-corrected value that is comparable across studies (Freedman *et al.* 2004).

We calculated λ as aggregated across three groups: linked, including only the autosome arm containing the causal SNP; unlinked, including the unlinked autosome that doesn't contain the causal SNP; and autosomal, for all sites across both autosomes two and three

Inflation factor measured across autosomes two and three was greatest for ILs, followed by 32-founder Hybrid Swarm, RILs, 128-founder Hybrid Swarm, and $_{50}$ Outbred populations. This order was observed whether the causal allele was common or rare, though with reduced values of λ at the lower allele frequency (Figure 7).

Only inbred populations displayed inflation on unlinked autosomes. When the causal allele is common (50% frequency), inflation on unlinked sites was greater for Inbred lines (median $\lambda = 1.17$, interquartile range or $IQR = 0.11$) than for RILs ($\lambda = 1.02, IQR = 0.07$). There was no inflation for unlinked chromosome in outbred populations, where $\lambda = 1.0$ with varying degrees of dispersion ($IQR = 0.10, 0.06$ and 0.03, respectively, for 32-founder HS, 128-founder HS, and F50 outbred populations). Unlinked sites remained inflated for ILs even when the causal allele was rare ($\lambda = 1.07, IQR = 0.09$). Distributions for λ across an extended range of autosome groups, PVE and allele frequencies are shown in Figure S11.

When we dissociated phenotype from genotype with purely random case-control assignment (i.e. PVE was set to 0% in our simulations), λ was centered at 1 (Figure S12). F_{50} outbred populations exhibited the lowest dispersion ($IQR = 0.02$), followed by 128-founder Hybrid Swarms ($IQR = 0.04$), RILs ($IQR = 0.06$), and 32-founder Hybrid Swarms or ILs ($IQR = 0.07$ each).

Frequency of sites segregating at appreciable frequency

The number of SNPs segregating amongst DGRP haplotypes with at least a given MAF strongly depends on the haplotype subset count for a given population (Figure 8). If only considering SNPs segregating at or above a frequency of 12.5% on chromosome arm 2L, a population founded by 8 lines will yield approximately twice as many SNPs

compared to a population founded by 128 lines (N=8 lines yields a median of 140K SNPs; N=128 lines yields a median of 71k SNPs). If the minimum MAF threshold is instead set to 5%, then populations with a greater number of lines exhibit a greater number of SNPs—with a maximum number of segregating sites with N=16 lines (median of 231.6k SNPs), nearly as many for 128 lines (median of 194k SNPs), and fewer for N=8 lines (median of 133k SNPs).

Discussion

Herein, we examined the feasibility and statistical properties of genome-wide association mapping using the Hybrid Swarm, an outbred population derived from limited and random outcrossing of an arbitrary number of founding strains. We show that it is possible to accurately reconstruct whole genomes from Hybrid Swarm populations using ultra-low coverage sequencing data (Figure 5). Genome-wide association mapping using the Hybrid Swarm approach performs as well as mapping in highly outbred F_{50} populations in a case-control GWAS framework (Figures 6, Supplemental Figure S10). While mapping using the Hybrid Swarm approach generally has reduced power compared to mapping using inbred lines (as would any outbred population in general) a limited number of generations of recombination reduces false positives arising from long-distance linkage disequilibrium present in founding strains (Figure 7, Supplemental Figure S11). Together, our results demonstrate the feasibility and potential of using the Hybrid Swarm approach for generating and genotyping outbred mapping populations in a cost-effective and computationally-efficient (Figure 1) manner.

Benefits of the Hybrid swarm Approach

The Hybrid Swarm approach is applicable to a wide variety of organisms and experimental designs, conferring potential benefits over inbred reference panels. These benefits are realized in three primary ways by: (1) allowing researchers to address questions that require heterozygotes; 2) reducing labor and the influence of cage-effects with random mating in a common environment; and 3) breaking down population structure when incorporating individuals from divergent populations. These benefits are

possible due to the ability to reconstruct genomes accurately and in a cost-effective manner for a large number of individuals.

Note that the Hybrid Swarm method is not limited to populations founded by inbred lines, as the technique can be applied to populations where phased genomes are available for all outbred founders. Research systems without inbred reference panels can thus make an up-front investment of fully phasing founder genomes to realize downstream savings of reconstructing progeny from low-coverage sequencing data. Due to the relative ease of generating phased genomes from a variety of long-read sequencing technologies (Pollard *et al.* 2018), the Hybrid Swarm method may enable association mapping in a wide variety of organisms.

Representation of heterozygotes

One clear difference between inbred and outbred mapping populations is the presence of heterozygotes. On the one hand, the presence of heterozygotes in outbred populations decreases power to detect association relative to inbred lines for an (semi-) additive allele with a given effect size (Figure 6, Supplemental Figure S10). However, the reduced statistical power of association mapping in outbred populations may be ameliorated by reduced inbreeding depression and by the ability to assess the heterozygous effects of alleles.

The ability to assess heterozygous effects of alleles will provide valuable insights into several interesting aspects of biology, such as the nature of dominance and the identity of regulatory polymorphisms. An increased understanding of dominance relationships and regulatory polymorphisms is important for advancing our understanding of quantitative trait variation and evolution. For instance, several theoretical models have shown that context dependent dominance of quantitative fitness traits can underlie the stable maintenance of polymorphisms subject to seasonally variable (Wittmann *et al.* 2017) or sexually antagonistic (Connallon and Chenoweth 2019) selection. The ability to efficiently map loci with context dependent dominance relationships will aid in the

652 understanding of the stability and abundance of polymorphisms maintained by these
653 forms of balancing selection.
654
655 Regulatory polymorphisms are known to underlie genetic variation in expression (Brem
656 *et al.* 2002; Cavet *et al.* 2003; Rockman and Kruglyak 2006) and this expression
657 variation can potentially be resolved to exact nucleotide differences (Grosveld *et al.*
658 1987; Rave-Harel *et al.* 1997; Bosma *et al.* 2002). The resulting differences in
659 expression can manifest as phenotypic changes to drive local adaptation (Kudaravalli
660 *et al.* 2009; Fraser *et al.* 2010; Fraser 2011, 2013). Allele-specific expression (ASE)
661 arising from cis-acting regulatory factors is a common mechanism to produce heritable
662 differences in expression (Yan *et al.* 2002; Cowles *et al.* 2002; Lo *et al.* 2003; Doss
663 2005). Because allelic expression biases are only produced (and detectable) in
664 heterozygotes, Hybrid Swarm populations facilitate the study of regulatory genetic
665 variation (i.e. ASE) as a driver of local adaptation in a variety of organisms.
666
667 *Undirected outbreeding in a common environment*
668 The Hybrid Swarm approach involves propagation of a single large outbred population
669 via undirected crossing. This design confers benefits over alternatives of either rearing
670 inbred lines separately or performing controlled crosses. First, a single population
671 reduces the influence of random block effects associated with rearing families or closely
672 related individuals in separate enclosures or defined areas. Second, random
673 outbreeding of a single population requires less labor compared to performing controlled
674 crosses or serial propagation of inbred lines. One drawback of the randomly outbred
675 method is susceptibility to loss of haplotypes due to genetic drift. The distribution of
676 haplotypes can also be skewed by line-specific differences in fitness or fecundity, with
677 such differences being observed for DGRP lines (Horváth and Kalinka 2016). To
678 attenuate haplotype dropout, it may be prudent to seed a Hybrid Swarm with a large
679 population of F1 hybrids produced by round-robin crosses. The F1 population would
680 then be followed by a limited number of generations (e.g., 4-5) of random outbreeding.
681

Hybrid Swarm breaks down population structure and linkage disequilibrium

Recombination between lines in the Hybrid Swarm approach allows for greater dissection of functional polymorphisms segregating between genetically structured populations. If an association study incorporates haplotypes from multiple distinct source populations, causal variants would segregate along with other linked variants. Thus, to identify genetic mechanisms of local adaptation and trait variation in general, it is necessary to minimize false positives from linked non-causal loci. Corrections due to relatedness can reduce the type I error rate to some degree (Yu *et al.* 2006; Price *et al.* 2010; Yang *et al.* 2014), and can be further reduced by a greater extent of recombination. Within mapping populations with many haplotypes such as the DGRP, long-distance linkage disequilibrium results from correlated occurrence of rare variants (Huang *et al.* 2014), potentially contributing to false positives in GWAS. This is reflected in our simulations by genome-wide inflation of λ, even across physically unlinked chromosomes, whereas five generations of recombination were sufficient to reduce this inflation (Figure 7).

Most notably, F_5 Hybrid Swarm populations performed equivalently to F_{50}outbred population in a case-control GWAS framework. This is likely owed to the large number of unique haplotypes within the Hybrid Swarm population, reducing the influence of long distance LD, and in turn reducing false positive GWAS hits. One interpretation is that only slightly recombinant populations comprised of a modest number of haplotypes are sufficient representations of highly outbred (or wild) populations in a GWAS framework. Inbred populations did exhibit greater power than outbred populations for identifying a causal locus, although this result is to be expected. Because we simulated a purely additive trait for which heterozygotes are equally likely to be assigned to either case or control group, heterozygotes contribute no statistical signal of association. Accordingly, for a causal allele segregating at 50% frequency, sample sizes for any outbred populations will be effectively half that of an inbred population.

The Hybrid Swarm method is similar but distinct from advanced intercross populations (AIPs), where AIPs result from crossing few lines (e.g., 8) for many generations

(Chesler 2014; Mackay and Huang 2018) and the Hybrid Swarm from crossing dozens to hundreds of lines for few generations. The choice to use an AIP or hybrid swarm population will influence the number of SNPs segregating at or above a desired minor allele frequency (Figure 8). For an association test to detect a causal variant with single-nucleotide precision, that variant must be segregating above a minor allele frequency required to detect phenotypic association at a given effect size and sample size. If sample size precludes sites segregating at a minor allele frequency below 1/8, then a population founded by 8 haplotypes would yield the greatest number of variants. If power is sufficient to detect association with alleles segregating above a frequency of 5%, then populations founded by 16+ lines would yield a greater number of variants (Figure 8). In cases where only few founding haplotypes are available, an AIP may be necessary, as the breakup of linkage disequilibrium can only be accomplished with many generations of crosses instead of leveraging greater haplotype diversity.

Computational Considerations

The simulations conducted for this analysis were made feasible by three primary innovations. First, the haplotype block file format allowed us to leverage information redundancy between related individuals and store highly compressed, lossless genotype information. With nearly 1/2000th the file size of a compressed VCF file, haplotype block files greatly reduced both the disk storage footprint and time required for disk write operations. Second, instead of performing forward-time simulations for every single iteration of GWAS, permuted subsets of simulated populations allowed for more rapid GWAS simulations. The format of haplotype block files facilitated permutations of the ancestry contained within a population's mosaic haplotypes, generating novel population genetic structure while preserving the forward-simulator's influence of drift and meiotic recombination (Figure 2). Third, instead of extracting site-specific genotypes for every individual, we decreased the number of computational operations by performing aggregate counts across all sites between adjacent recombination events in the population (Figure 3).

743 Importantly, selecting a subset of most-likely-ancestors results in maximum
744 computational complexity that remains constant with increasing number of founding
745 lines, instead of complexity increasing at a greater-than-linear rate. This means that the
746 larger the pool of unique haplotypes that an individual descends from, the greater
747 speedup of our pipeline relative to other methods. Although computational speed has
748 been shown to be reduced by haplotype pre-phasing (Howie *et al.* 2011), to our
749 knowledge, pre-phasing has not been demonstrated with ultra-low coverage sequencing
750 on the order of 0.005-0.05X. As a result, pre-phasing would likely require greater
751 sequencing effort, negating the benefit of low coverage reconstruction. Computational
752 search space can also be reduced if an individual's pedigree is known with certainty,
753 however controlled crosses can be laborious, and may lead to cage-specific effects.
754
755
756 **Applying the Hybrid Swarm approach**
757 At minimum, the Hybrid Swarm approach requires a sequenced set of individuals for
758 founding a recombinant population. Although our simulations presented here were
759 conducted with inbred founding lines, genome reconstructions can similarly be
760 performed with any phased genomes. For example, 16 phased outbred founders could
761 be treated as 32 independent haplotypes. Phased genomes are becoming increasingly
762 accessible with the advent of long-read sequencing platforms and phasing software
763 (Chin *et al.* 2016; Mostovoy *et al.* 2016; Seo *et al.* 2016), allowing this technique to be
764 applied to even more systems. Optionally, a recombination rate map for the population
765 can be provided, otherwise recombination is assumed to occur with equal user-defined
766 probability across any chromosome.
767
768 As a first step, power analyses using our rapid association test simulation pipeline will
769 inform choices of sample size and mapping population design (Figure 3). After
770 determining a feasible sample size for a given SNP of minimum percent variation
771 explained, researchers can evaluate the accuracy of low-coverage chromosome
772 reconstructions for a simulated proposed mapping population. Note that while we
773 performed association tests in a case/control framework, the relative power of the

Hybrid Swarm is expected to be the same for quantitative traits, which could garner additional power from sampling individuals from phenotypic extremes (D. Li, Lewinger, Gauderman, Murcray, & Conti, 2011).

For our simulations, we parameterized chromosome reconstructions using a maximum of $N = 16$ most-likely ancestors (MLAs) and $S = 5000$ SNPs, which required less than 3 GB of memory and completed in under 5 minutes on a single core. However, these values may not be ideal for all systems. It may be necessary to select greater number of MLAs prior to reconstruction if haplotypes are difficult to differentiate due to being less divergent (i.e. exhibiting lower θ_π) than those simulated here. For example, reconstruction accuracy was low for coalescent-derived mapping populations modeled with $\theta = 4 \times 10^{-5}$ (Supplemental Figure S8), which may reflect those of *C. elegans* (Barriere and Félix 2005). Further, 5000 SNPs may be an over- or under-estimate of those required in other systems. Because recombination between haplotypes can only be inferred at sampled variable sites, SNP density directly influences how close inferred breakpoints will be resolved with respect to their actual position. The models described in Figure 1 can be used to estimate the memory and runtime required for a given number of input ancestor haplotypes.

To evaluate whether low coverage sequencing data will yield accurate genotype estimates for a given proposed mapping population, researchers can test reconstruction accuracy *in silico*. We provide a convenient forward-simulation R script for this purpose that generates output in the haplotype map format (Figure 2). Simulated individuals can then be ran through the simulated sequencing and mapping pipeline at a desired level of coverage. After generating simulated mapped individuals, researchers can optimize the number of MLAs and HARP threshold that provide most effective MLA selection for their mapping population (Figure 4). This step may reveal haplotypes that are consistently problematic or inaccurately chosen, which can be excluded from further simulations (and when generating the true mapping population). Researchers can then perform chromosome reconstruction using the optimized MLA selection parameters and evaluate whether accuracy is acceptable (Figure 5).

After performing chromosome reconstructions, a quality control step may be applied whereby troublesome regions are masked. For example, a reconstructed chromosome with a sequence of short recombination blocks could be masked prior to evaluating genotyping accuracy or performing association testing. In our simulations, it was surprisingly difficult to diagnose exact factors contributing to the least accurate reconstructions. However, these highly recombinant reconstructions still achieved 90-99% accuracy, suggesting that accuracy may be achieved even for anomalous hyper-recombinant individuals (Figure 5). Optimized parameters can then be applied to a genuine mapping population akin to the simulated one.

Conclusions

An outbred high-resolution mapping population that can be generated in little time is an attractive option for researchers, but such mapping populations have been prohibited by genotyping costs or computational requirements to impute genotypes from ultra-low sequencing data. Our work demonstrates the feasibility of the Hybrid Swarm as a cost-effective method of fine-scale genetic mapping in an outbred population and provides a computationally efficient framework for GWAS power analysis.

References

Barrière, A. and Félix, M.-A. Natural variation and population genetics of *Caenorhabditis elegans* (December 26, 2005), *WormBook*, ed. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.43.1, http://www.wormbook.org.

Bergland A. O., H. Chae, Y.-J. Kim, and M. Tatar, 2012 Fine-Scale Mapping of Natural Variation in Fly Fecundity Identifies Neuronal Domain of Expression and Function of an Aquaporin. PLoS Genet. 8: e1002631.

Bosma P. J., J. R. Chowdhury, C. Bakker, S. Gantla, A. de Boer, *et al.*, 2002 The Genetic Basis of the Reduced Expression of Bilirubin UDP-Glucuronosyltransferase 1 in Gilbert's Syndrome. N. Engl. J. Med. 333: 1171–1175.

Brem R. B., G. Yvert, R. Clinton, and L. Kruglyak, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science 296: 752–5.

Broad Institute, 2015a The Picard toolkit. https://broadinstitute.github.io/picard/.

839 Broad Institute, 2015b Genome Analysis Toolkit: Variant Discovery in High-Throughput
840 Sequencing Data. https://software.broadinstitute.org/gatk/.

841 Cavet G., P. S. Linsley, M. Mao, R. B. Stoughton, S. H. Friend, *et al.*, 2003 Genetics of
842 gene expression surveyed in maize, mouse and man. Nature 422: 297–302.

843 Cheng R., J. E. Lim, K. E. Samocha, G. Sokoloff, M. Abney, *et al.*, 2010 Genome-wide
844 association studies and the problem of relatedness among advanced intercross
845 lines and other highly recombinant populations. Genetics 185: 1033–1044.

846 Chesler E. J., D. R. Miller, L. R. Branstetter, L. D. Galloway, B. L. Jackson, *et al.*, 2008
847 The Collaborative Cross at Oak Ridge National Laboratory: Developing a powerful
848 resource for systems genetics. Mamm. Genome 19: 382–389.

849 Chesler E. J., 2014 Out of the bottleneck: the Diversity Outcross and Collaborative
850 Cross mouse populations in behavioral genetics research. Mamm. Genome 25: 3–
851 11.

852 Chia R., F. Achilli, M. F. W. Festing, and E. M. C. Fisher, 2005 The origins and uses of
853 mouse outbred stocks. Nat. Genet. 37: 1181–1186.

854 Chin C.-S., P. Peluso, F. J. Sedlazeck, M. Nattestad, G. T. Concepcion, *et al.*, 2016
855 Phased diploid genome assembly with single-molecule real-time sequencing. Nat.
856 Methods 13: 1050–1054.

857 Comeron J. M., R. Ratnappan, and S. Bailin, 2012 The Many Landscapes of
858 Recombination in Drosophila melanogaster. PLoS Genet. 8: 33–35.

859 Connallon T., and S. F. Chenoweth, 2019 Dominance reversals and the maintenance of
860 genetic variation for fitness. PLoS Biol. 17: 1–11.

861 Cowles C. R., J. N. Hirschhorn, D. Altshuler, and E. S. Lander, 2002 Detection of
862 regulatory variation in mouse genes. Nat. Genet. 32: 432–437.

863 Danecek P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, *et al.*, 2011 The variant call
864 format and VCFtools. Bioinformatics 27: 2156–2158.

865 Doss S., 2005 Cis-acting expression quantitative trait loci in mice. Genome Res. 15:
866 681–691.

867 Fraser H. B., A. M. Moses, and E. E. Schadt, 2010 Evidence for widespread adaptive
868 evolution of gene expression in budding yeast. Proc. Natl. Acad. Sci. 107: 2977–
869 2982.

870 Fraser H. B., 2011 Genome-wide approaches to the study of adaptive gene expression
871 evolution. BioEssays 33: 469–477.

872 Fraser H. B., 2013 Gene expression drives local adaptation in humans. Genome Res.

873 23: 1089–1096.

874 Freedman M. L., D. Reich, K. L. Penney, G. J. McDonald, A. A. Mignault, *et al.*, 2004
875 Assessing the impact of population stratification on genetic association studies.
876 Nat. Genet. 36: 388–393.

877 Grosveld F., G. B. van Assendelft, D. R. Greaves, and G. Kollias, 1987 Position-
878 independent, high-level expression of the human beta-globin gene in transgenic
879 mice. Cell 51: 975–85.

880 Horváth B., and A. T. Kalinka, 2016 Effects of larval crowding on quantitative variation
881 for development time and viability in Drosophila melanogaster. Ecol. Evol. 6: 8460–
882 8473.

883 Howie B., J. Marchini, and M. Stephens, 2011 Genotype imputation with thousands of
884 genomes. G3 1: 457–70.

885 Huang X., M.-J. Paulo, M. Boer, S. Effgen, P. Keizer, *et al.*, 2011 Analysis of natural
886 allelic variation in Arabidopsis using a multiparent recombinant inbred line
887 population. Proc. Natl. Acad. Sci. U. S. A. 108: 4488–93.

888 Huang B. E., A. W. George, K. L. Forrest, A. Kilian, M. J. Hayden, *et al.*, 2012 A
889 multiparent advanced generation inter-cross population for genetic analysis in
890 wheat. Plant Biotechnol. J. 10: 826–839.

891 Huang W., A. Massouras, Y. Inoue, J. Peiffer, M. Ramia, *et al.*, 2014 Natural variation in
892 genome architecture among 205 Drosophila melanogaster Genetic Reference
893 Panel lines. Genome Res. 24: 1193–1208.

894 Kessner D., T. L. Turner, and J. Novembre, 2013 Maximum likelihood estimation of
895 frequencies of known haplotypes from pooled sequence data. Mol. Biol. Evol. 30:
896 1145–58.

897 King E. G., S. J. Macdonald, and A. D. Long, 2012a Properties and power of the
898 Drosophila synthetic population resource for the routine dissection of complex
899 traits. Genetics 191: 935–949.

900 King E. G., C. M. Merkes, C. L. McNeil, S. R. Hoofer, S. Sen, *et al.*, 2012b Genetic
901 dissection of a model complex trait using the Drosophila Synthetic Population
902 Resource. Genome Res. 22: 1558–1566.

903 Kover P. X., W. Valdar, J. Trakalo, N. Scarcelli, I. M. Ehrenreich, *et al.*, 2009 A
904 Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in
905 Arabidopsis thaliana, (R. Mauricio, Ed.). PLoS Genet. 5: e1000551.

906 Krämer N., N. Ranc, N. Meyer, M. Ouzunova, C. Lehermeier, *et al.*, 2014 Usefulness of
907 Multiparental Populations of Maize (Zea mays L.) for Genome-Based Prediction .
908 Genetics 198: 3–16.

909 Kudaravalli S., J. B. Veyrieras, B. E. Stranger, E. T. Dermitzakis, and J. K. Pritchard,
910 2009 Gene expression levels are a target of recent natural selection in the human
911 genome. Mol. Biol. Evol. 26: 649–658.

912 Lack J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, *et al.*, 2015
913 The drosophila genome nexus: A population genomic resource of 623 Drosophila
914 melanogaster genomes, including 197 from a single ancestral range population.
915 Genetics 199: 1229–1241.

916 Li R., M. A. Lyons, H. Wittenburg, B. Paigen, and G. A. Churchill, 2005 Combining data
917 from multiple inbred line crosses improves the power and resolution of quantitative
918 trait loci mapping. Genetics 169: 1699–709.

919 Li H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, *et al.*, 2009 The Sequence
920 Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

921 Li H., 2011 wgsim (short read simulator). https://github.com/lh3/wgsim.

922 Li H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-
923 MEM. arXiv:1303.3997.

924 Lo H. S., Z. Wang, Y. Hu, H. H. Yang, S. Gere, *et al.*, 2003 Allelic variation in gene
925 expression is common in the human genome. Genome Res. 13: 1855–62.

926 Mackay T. F. C., and W. Huang, 2018 Charting the genotype–phenotype map: lessons
927 from the Drosophila melanogaster Genetic Reference Panel. Wiley Interdiscip. Rev.
928 Dev. Biol. 7: e289.

929 MacKay T. F. C. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, *et al.*, 2012
930 The Drosophila melanogaster Genetic Reference Panel. Nature 482: 173–178.

931 Monir M. M., and J. Zhu, 2017 Comparing GWAS Results of Complex Traits Using Full
932 Genetic Model and Additive Models for Revealing Genetic Architecture. Sci. Rep. 7:
933 38600.

934 Mostovoy Y., M. Levy-Sakin, J. Lam, E. T. Lam, A. R. Hastie, *et al.*, 2016 A hybrid
935 approach for de novo human genome sequence assembly and phasing. Nat.
936 Methods 13: 587–590.

937 Mott R., C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint, 2000 A method for fine
938 mapping quantitative trait loci in outbred animal stocks. Proc. Natl. Acad. Sci. 97:
939 12649–54.

940 Nikpay M., A. Goel, H. H. Won, L. M. Hall, C. Willenborg, *et al.*, 2015 A comprehensive
941 1000 Genomes-based genome-wide association meta-analysis of coronary artery
942 disease. Nat. Genet. 47: 1121–1130.

943 Paul R. Staab, Sha Zhu, Dirk Metzler, and Gerton Lunter, 2015 {scrm}: efficiently

944 simulating long sequences using the approximated coalescent with recombination.
945 Bioinformatics 31: 1680–1682.

946 Pollard M. O., D. Gurdasani, A. J. Mentzer, T. Porter, and M. S. Sandhu, 2018 Long
947 reads: their purpose and place. Hum. Mol. Genet. 27: R234–R241.

948 Price A. L., N. A. Zaitlen, D. Reich, and N. Patterson, 2010 New approaches to
949 population stratification in genome-wide association studies. Nat. Rev. Genet. 11:
950 459–463.

951 R Core Team, 2016 R: A Language and Environment for Statistical Computing

952 Rave-Harel N., E. Kerem, M. Nissim-Rafinia, I. Madjar, R. Goshen, *et al.*, 1997 The
953 molecular basis of partial penetrance of splicing mutations in cystic fibrosis. Am. J.
954 Hum. Genet. 60: 87–94.

955 Rockman M. V., and L. Kruglyak, 2006 Genetics of global gene expression. Nat. Rev.
956 Genet. 7: 862–872.

957 Seo J. S., A. Rhie, J. Kim, S. Lee, M. H. Sohn, *et al.*, 2016 De novo assembly and
958 phasing of a Korean human genome. Nature 538: 243–247.

959 Singh R., I. T. Lobina, M. Thomson, S. McCouch, C. Dilla-Ermita, *et al.*, 2013 Multi-
960 parent advanced generation inter-cross (MAGIC) populations in rice: progress and
961 potential for genetics research and breeding. Rice 6: 11.

962 Spencer C. C. A., Z. Su, P. Donnelly, and J. Marchini, 2009 Designing genome-wide
963 association studies: Sample size, power, imputation, and the choice of genotyping
964 chip. PLoS Genet. 5.

965 Srivastava A., A. P. Morgan, M. L. Najarian, V. K. Sarsani, J. S. Sigmon, *et al.*, 2017
966 Genomes of the Mouse Collaborative Cross. Genetics 206: 537–556.

967 Stevenson M., 2018 epiR: Tools for the Analysis of Epidemiological Data.
968 https://rdrr.io/cran/epiR/.

969 Wittmann M. J., A. O. Bergland, M. W. Feldman, P. S. Schmidt, and D. A. Petrov, 2017
970 Seasonally fluctuating selection can maintain polymorphism at many loci via
971 segregation lift. Proc. Natl. Acad. Sci. 114: E9932–E9941.

972 Wu Y., Z. Zheng, P. M. Visscher, and J. Yang, 2017 Quantifying the mapping precision
973 of genome-wide association studies using whole-genome sequencing data.
974 Genome Biol. 18: 1–10.

975 Yan H., W. Yuan, V. E. Velculescu, B. Vogelstein, and K. W. Kinzler, 2002 Allelic
976 variation in human gene expression. Science 297: 1143.

977 Yang J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price, 2014 Advantages

and pitfalls in the application of mixed-model association methods. Nat. Genet. 46: 100–106.

Yang C., Y. Wang, W. Xu, Z. Liu, S. Zhou, *et al.*, 2018 Genome-wide association study using diversity outcross mice identified candidate genes of pancreatic cancer. Genomics 0–1.

Yu J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203–208.

Zhang J., K. Kobert, T. Flouri, and A. Stamatakis, 2014 PEAR: A fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30: 614–620.

Zheng C., M. P. Boer, and F. A. van Eeuwijk, 2015 Reconstruction of genome ancestry blocks in multiparental populations. Genetics 200: 1073–1087.

Zheng C., M. P. Boer, and F. A. van Eeuwijk, 2018 Accurate genotype imputation in multiparental populations from low-coverage sequence. Genetics 210: 71–82.

Figures

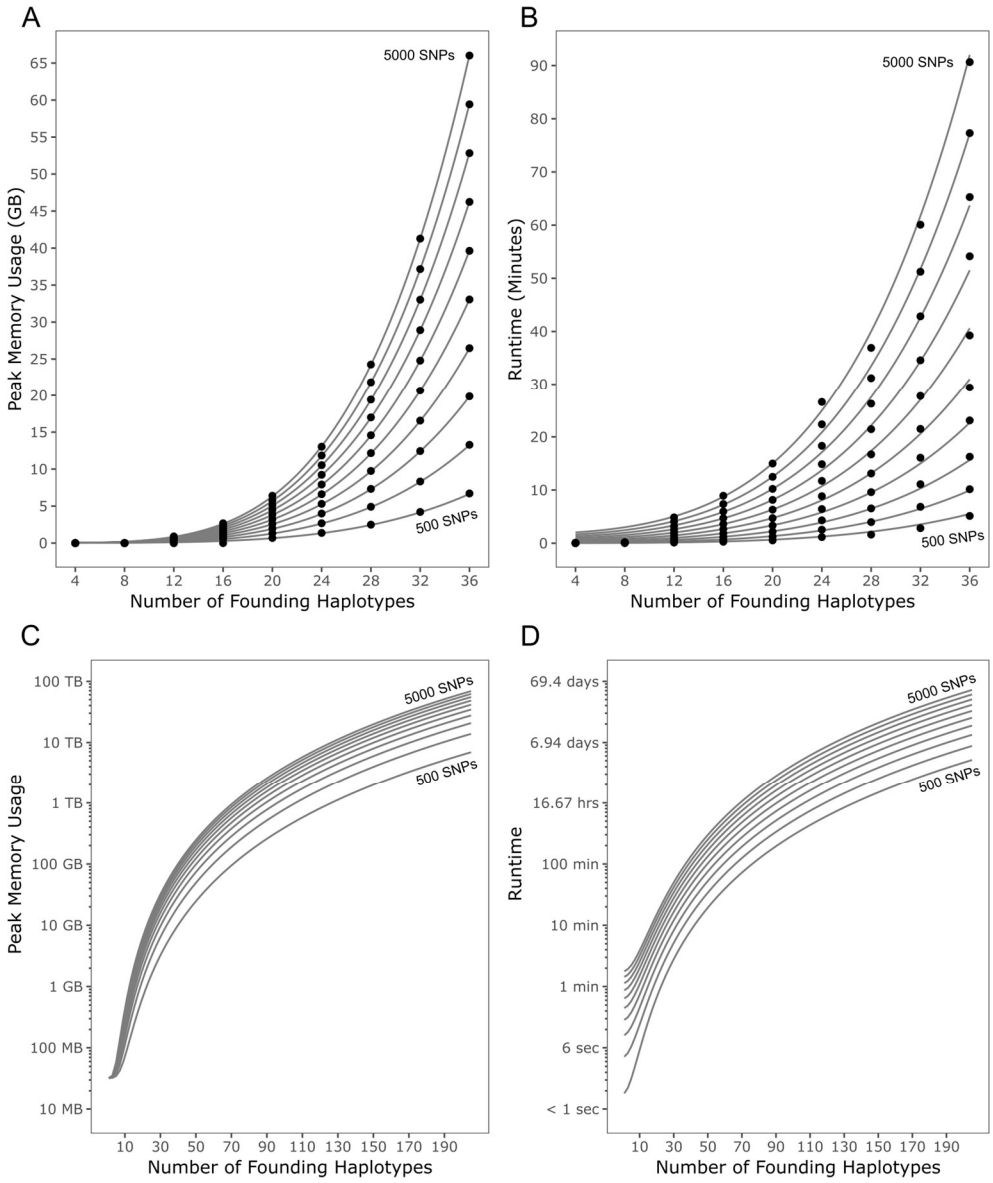

Figure 1. Resource usage of RABBIT during haplotype reconstruction. All reconstructions involve the same simulated 2L chromosome arm comprised of four haplotypes. Simulations included varied numbers of founding haplotypes (*N*) and a randomly selected set of markers (number of SNPs, *S*, incremented in steps of 500). All simulations included, at minimum, the four true haplotypes for the simulated individual. In **A** and **B**, points depict the mean of empirical values (over 10 replicates) and gray lines depict the defined regression models. Predicted peak memory usage and runtime are displayed on a log scale over a greater range for number of founding haplotypes in **C** and **D**, respectively.

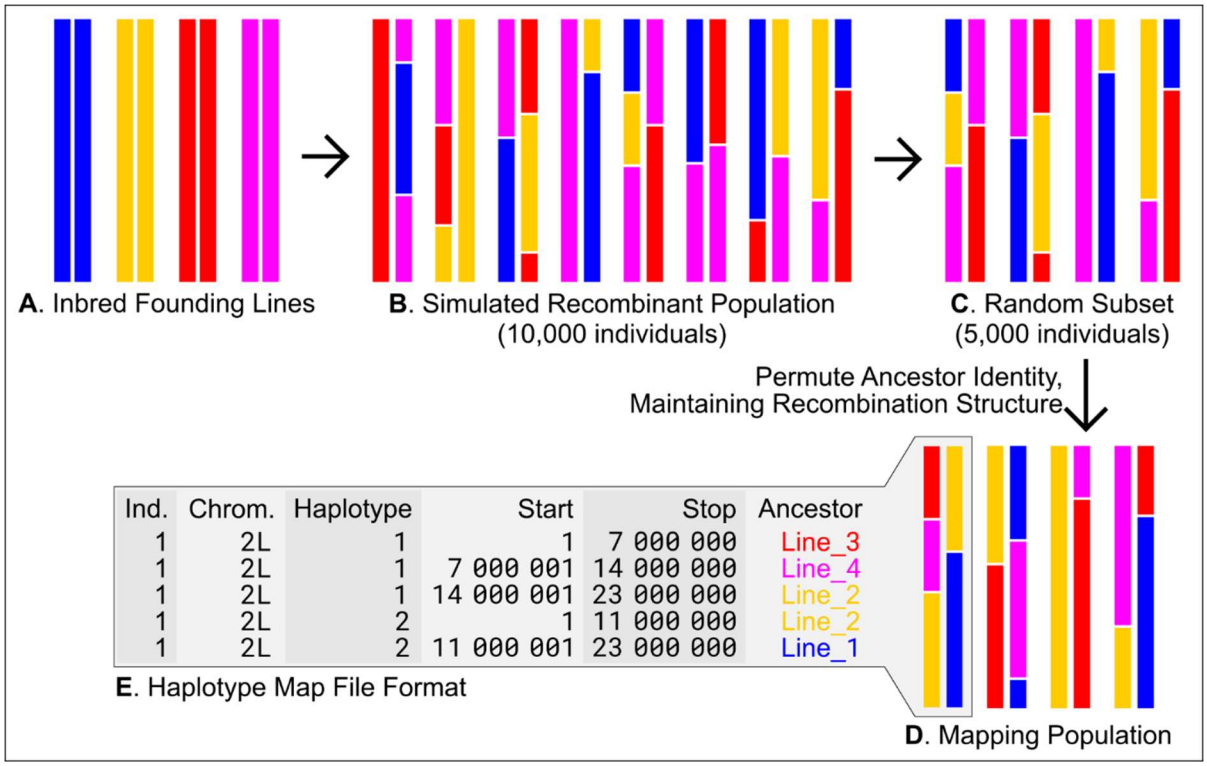

Figure 2. Basic structure of the forward simulator pipeline. Inbred founding lines (**A**) are randomly intercrossed to produce a recombinant population (**B**). Rapid generation of independent mapping populations is achieved by random down-sampling (**C**) and permutation of ancestry (**D**). Population genetic data is encoded in a highly compressed format **(E)** that references the positions of haplotype blocks instead of genotypes at every site, enabling us to generate 500 mapping populations for a given parameter combination. Individuals are probabilistically assigned to case or control groups based on genotype at a randomly chosen causal SNP segregating at a specified frequency.

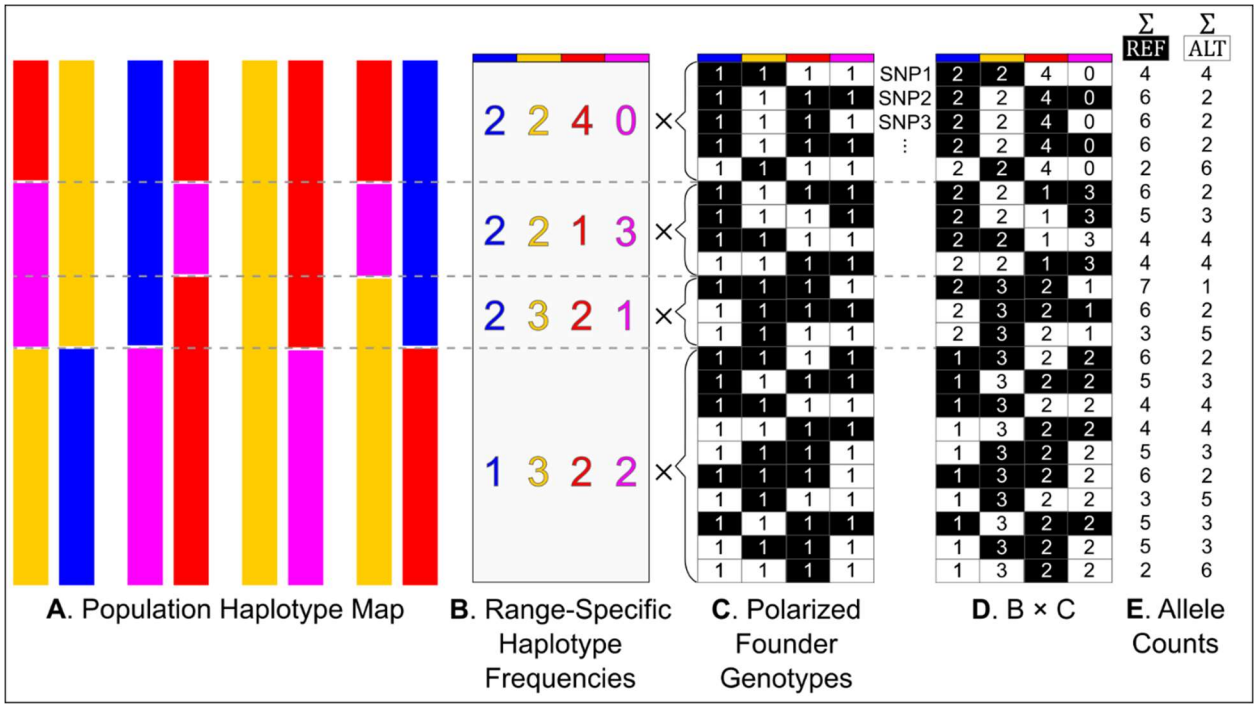

Figure 3. Schematic for rapid association testing with the haplotype block files. For a given population represented by a haplotype map file (**A**), all SNPs between sorted breakpoints (indicated by dashed lines) will share identical aggregated haplotype frequencies (**B**). Haplotype frequencies are multiplied by a founder genotype matrix (**C**) where alleles are coded reference (black cells) and alternate (white cells). Conditional row sums of the resulting matrix (**D**) yields reference and alternate frequencies at each locus (**E**), to be used for χ^2 tests of independence.

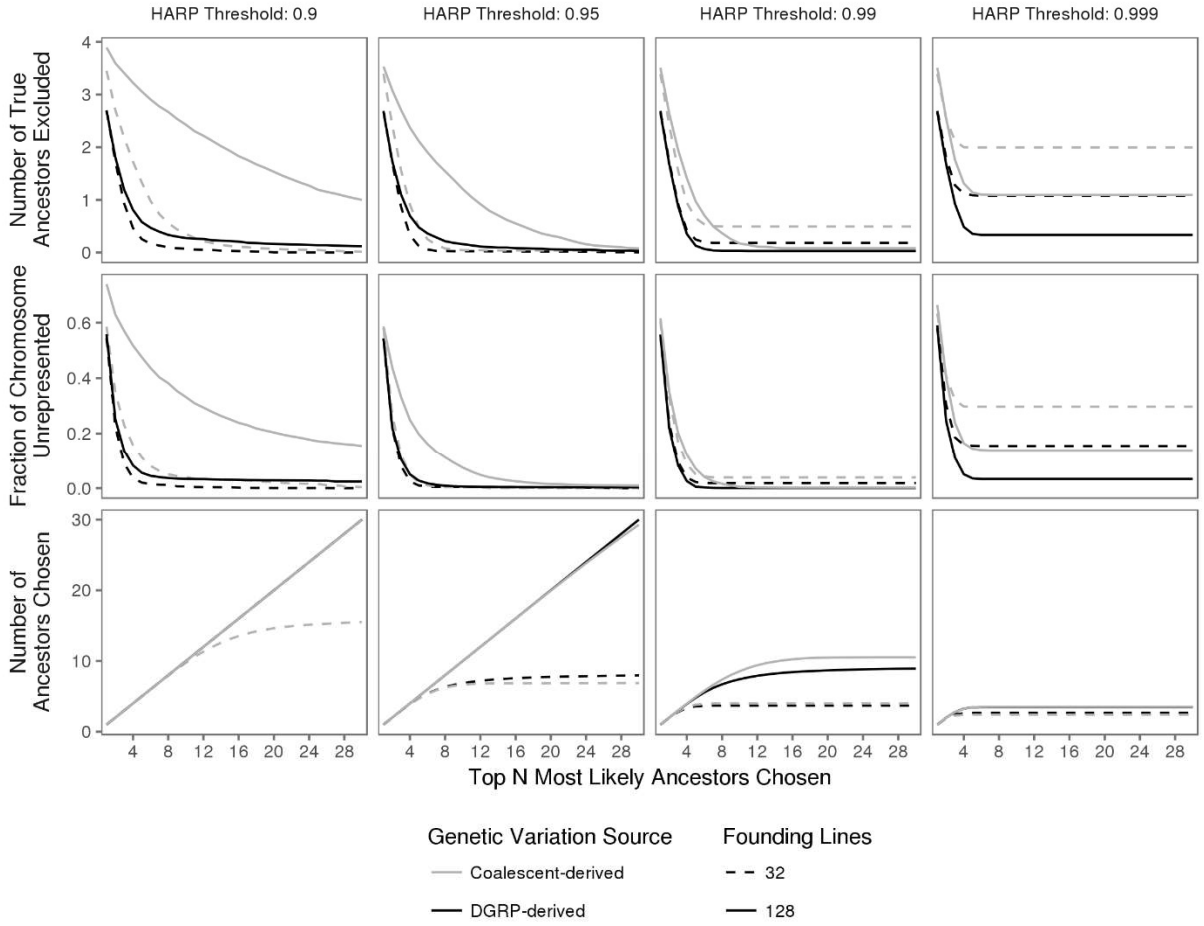

Figure 4. Optimization curves for Most-Likely-Ancestor (MLA) selection. Increasing the upper limit for the number of MLAs chosen reduces the number of true ancestors missed, similarly reducing the fraction of a given chromosome that is not represented within the selected set of MLAs. Ancestors that fail to pass the HARP threshold across all genomic windows are not selected, resulting in realized sets of MLAs (Number of Ancestors Chosen) below the upper-limit allowed (x-axis). Data shown reports means across 400 replicates made up of 100 simulated individuals (4 autosomes each for coalescent simulations, 4 autosome arms each for DGRP simulations) per parameter combination. Coalescent-derived populations described here were simulated with $N_e = 10^6$ and $\mu = 5 \times 10^{-9}$.

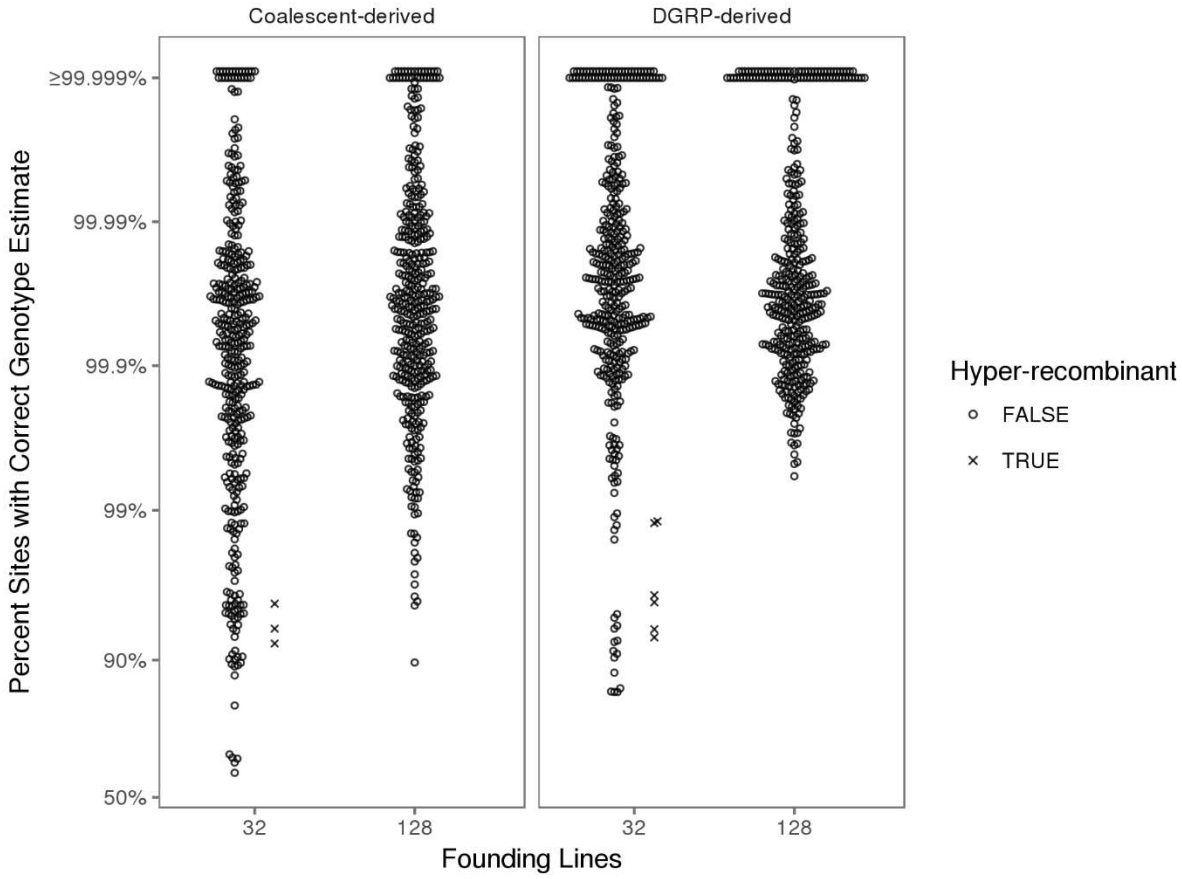

Figure 5. Accuracy of genome reconstruction pipeline for simulated F_5 Hybrid Swarm individuals.
Reconstructions were performed for populations simulated as being founded by either 32 or 128 inbred lines at 0.05X sequencing coverage with up to 16 MLAs as determined with a HARP threshold of 0.99. Accuracy, calculated as the per-chromosome fraction of variable sites with a correct diploid genotype estimate, is shown on logit-transformed scale. Values are coded depending on the number of estimated recombination events, with highly recombinant estimates ($\geq$10 recombination events) displayed as an $\times$. Each parameter combination includes 400 reconstructed autosomes (individual circles) for 100 simulated individuals. The coalescent-derived individuals displayed here were simulated with an effective population size of $N_e = 1 \times 10^6$ and mutation rate $\mu = 5 \times 10^{-9}$.

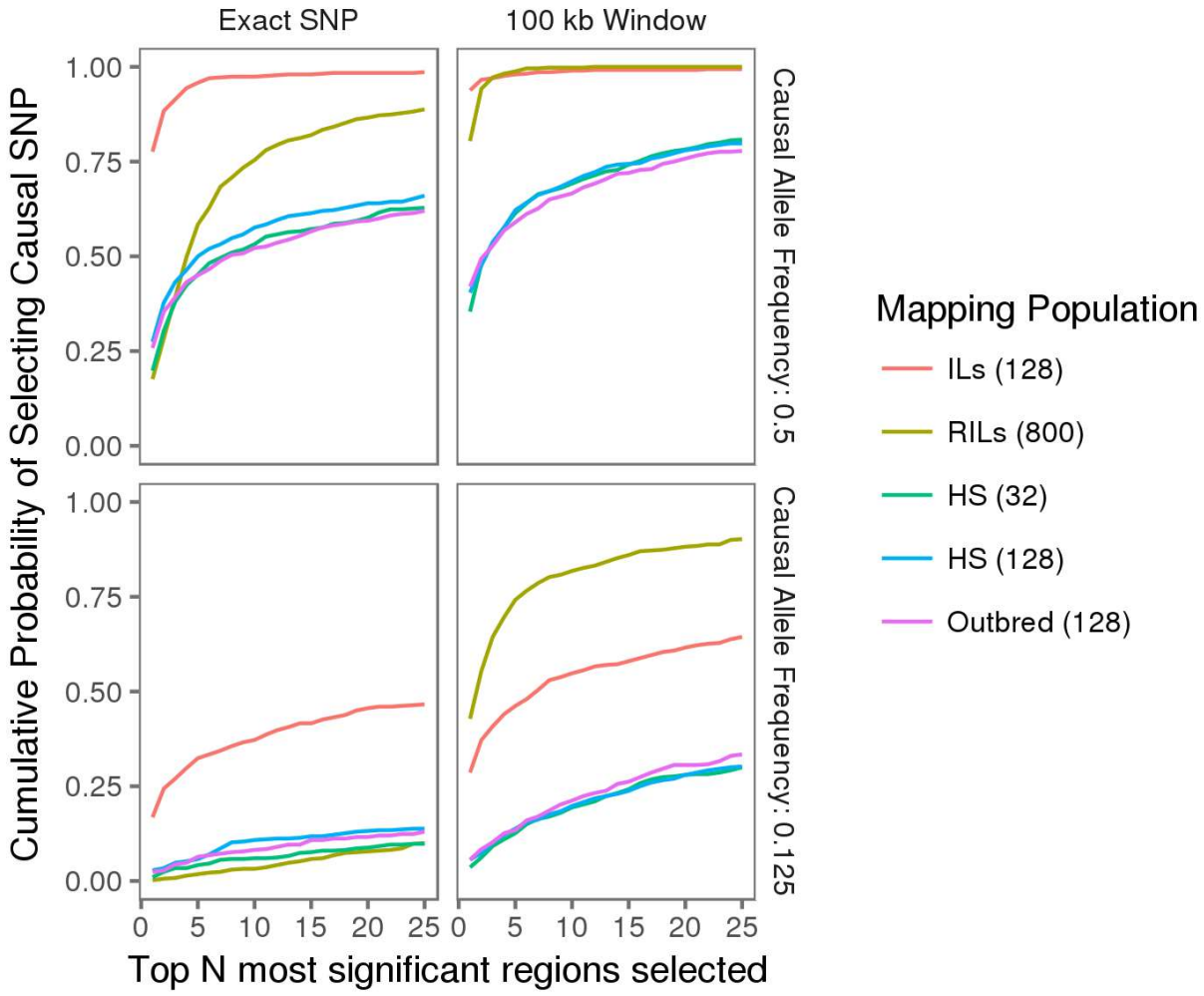

Figure 6. Accuracy of simulated GWAS for various mapping populations. Plots display the cumulative probability of including a causal SNP when selecting the top N most significant SNPs, or 100kb windows around those SNPs, out of 500 simulated GWAS (each comprised of 5000 individuals phenotypically assigned in a case-control framework). Homozygotes for the reference allele were assigned to the case group with 45% probability, while homozygotes for the alternate allele were assigned to the case group with 55% probability (a difference of 10%), and heterozygotes are assigned to case and control groups with equal probability. **ILs**: inbred lines. **RILs**: recombinant inbred lines. **HS**: Hybrid Swarm populations founded by 32 or 128 lines. **Outbred**: An F_{50} population founded by 128 lines.

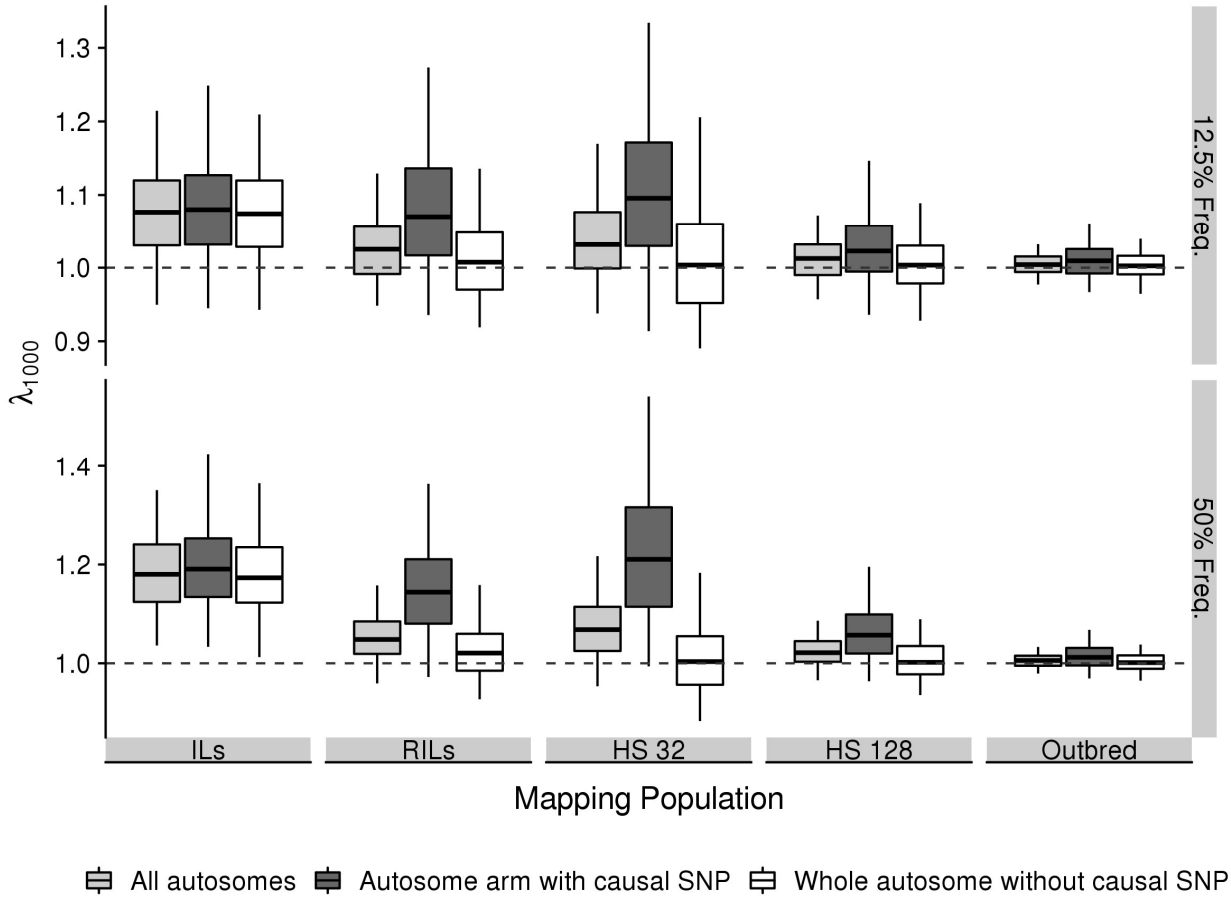

Figure 7. Genomic Inflation Factor (GIF, λ_{1000}) for simulated GWAS with a causal allele segregating at a specified frequency.
GIF is calculated genome-wide (across all autosomes); on the autosome arm containing the causal allele (linked); and for sites on the autosome physically unlinked to the causal allele. λ is calculated as the ratio of observed to expected χ^2 values, and a correction is performed to produce the null expectation given the sample size had actually been 1000 individuals (see Materials and Methods for details). Data are averaged over 500 simulated GWAS (each comprised of 5000 individuals phenotypically assigned in a case-control framework). Homozygotes for the reference allele were assigned to the case group with 45% probability, while homozygotes for the alternate allele were assigned to the case group with 55% probability (a difference of 10%), and heterozygotes are assigned to case and control groups with equal probability. Boxes represent the median and interquartile range; whiskers extending to the lower and upper bounds of the 95% quantiles. **ILs**: 128 Inbred Lines. **RILs**: 800 Recombinant Inbred Lines. **HS**: Hybrid Swarm with 32 or 128 founding lines. **Outbred**: F_{50} population founded by 128 inbred lines.

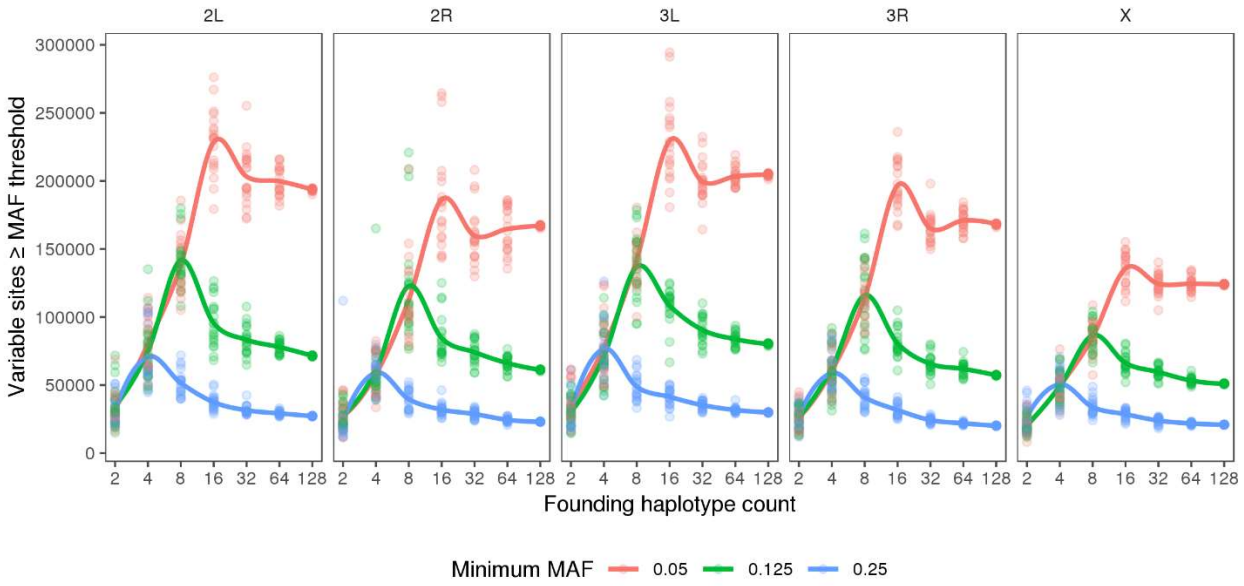

Figure 8. Counts of variable sites depending on number of founding DGRP haplotypes Each point represents the number of sites segregating at or above a given minor allele frequency threshold when drawing *N* haplotypes, with 20 replicates per parameter combination. With a minimum minor allele frequency (MAF) of 12.5%, a population founded by eight haplotypes exhibits approximately double the number of variable sites compared to a population founded by 128 haplotypes. With a minimum MAF of 5%, populations with eight founding haplotypes present with fewer SNPs compared to populations founded by 16 or more haplotypes.

Population	N Founders	ρ	$\overline{\Delta}$	σ_Δ
DGRP	128	0.986	-0.015	0.25
DGRP	32	0.502	0.17	2.15
Coalescent	128	0.956	-0.17	0.44
Coalescent	32	0.759	-0.31	1.26

Table 1. Accuracy of estimated number of recombination events following chromosome reconstruction.
A high concordance correlation coefficient (Lin's ρ) indicates agreement between estimated and true recombination counts for 400 reconstructed chromosomes (coalescent-derived populations) or chromosome arms (DGRP-derived populations). Coalescent-derived populations are described across a range of values for effective population size N_e and mutation rate μ. $\overline{\Delta}$ and σ_Δ denote mean and standard deviation, respectively, of difference between estimated and true recombination counts. Reconstructions were performed with a maximum of 16 most-likely-ancestors with a HARP threshold of 0.99 (see methods for more details).

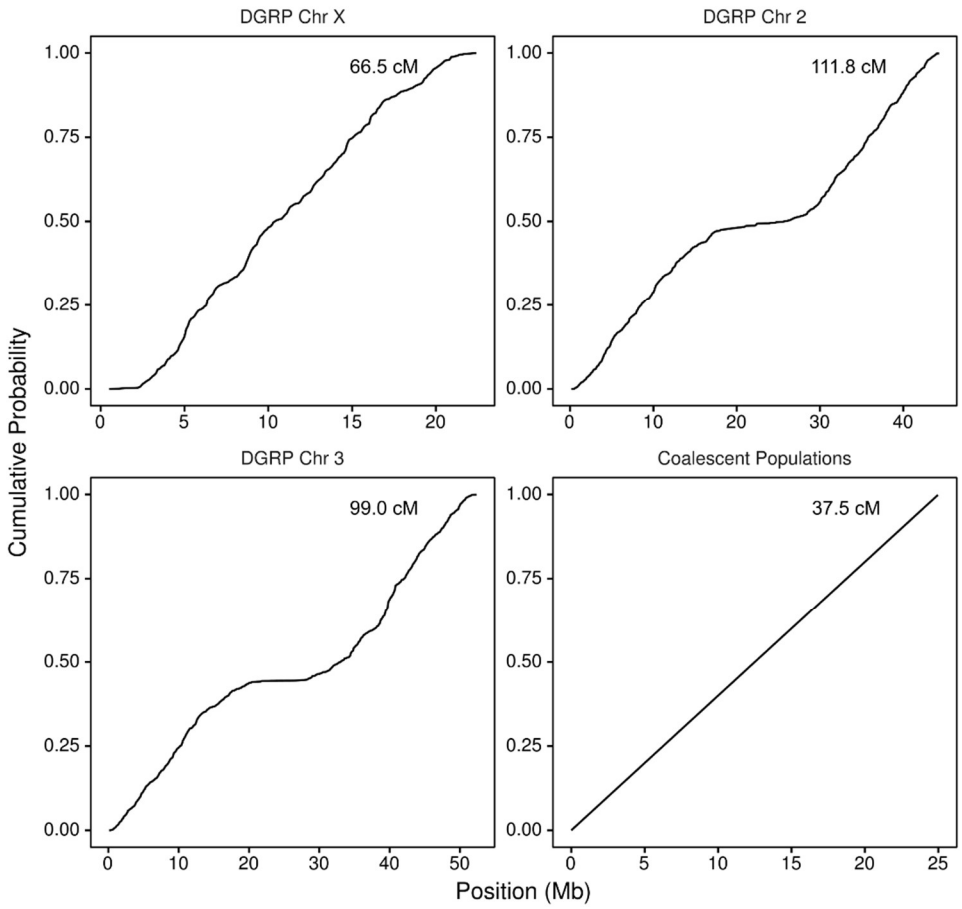

Figure S1. Recombination probability functions used for simulated individuals. Recombination is modeled as a Poisson process, with position sampled from linear interpolation of recombination rates measured in *Drosophila melanogaster* by Comeron et al. (2012). The frequency of recombination samples cumulative map distance (inset, e.g. a 99 cM chromosome is modeled as a Poisson variable with an expected value of $\lambda = 0.99$). For DGRP-derived individuals, recombination was simulated for full chromosomes two and three, and reconstructions were then conducted independently for arms 2L, 2R, 3L, and 3R.

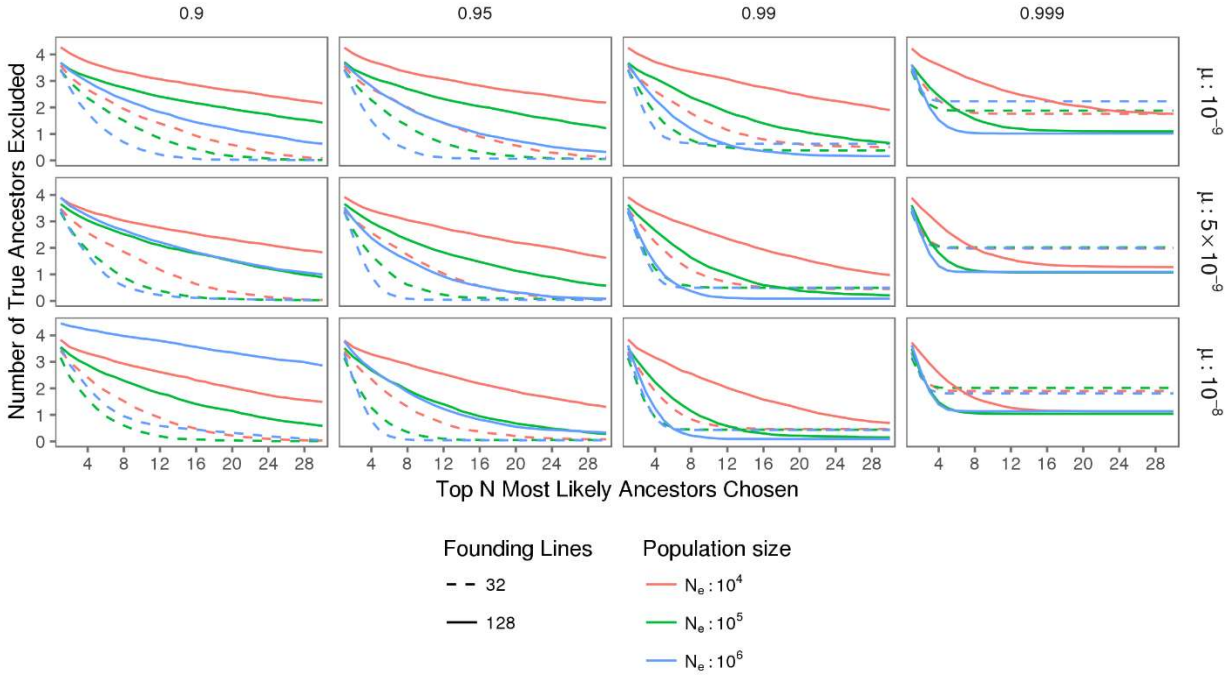

Figure S2. Optimization curves for Most-Likely-Ancestor inclusion, by count, in SCRM-derived F_5 hybrid swarm individuals

The number of missed true ancestors is shown as a function of the number of ancestors chosen across a range of HARP threshold values (0.9 to 0.999), effective population sizes (N_e) and mutation rates (μ). Values are averaged across 400 reconstructed autosomes (from 100 individuals) per parameter combination.

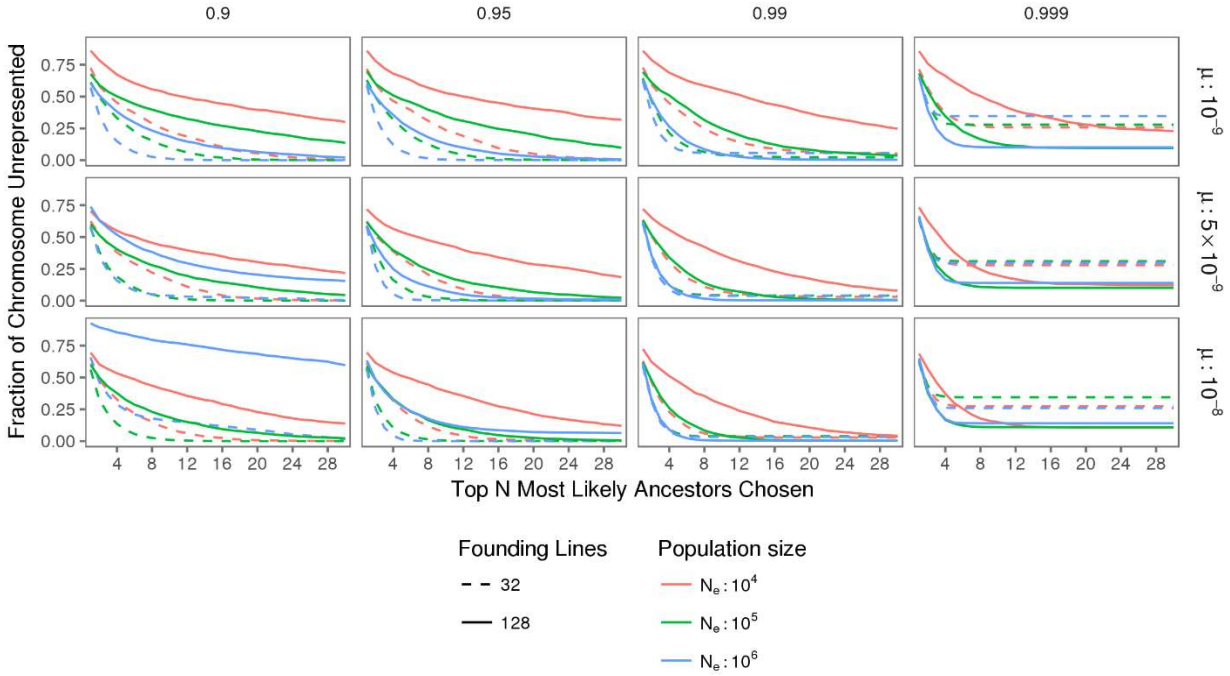

Figure S3. Optimization curves for Most-Likely-Ancestor inclusion, by chromosome representation, in simulated SCRM-derived F_5 hybrid swarm individuals. The proportion of the chromosome not covered by the chosen ancestors is shown as a function of the number of ancestors chosen for populations founded by either 32 or 128 inbred founding lines across a range of HARP threshold values (0.9 to 0.999), effective population sizes (N_e) and mutation rates (μ). Each line summarizes the arithmetic mean fraction of sites where the true ancestor is not included within the inferred set of Most-Likely-Ancestors. Values are averaged across 400 reconstructed autosomes (from 100 individuals) per parameter combination.

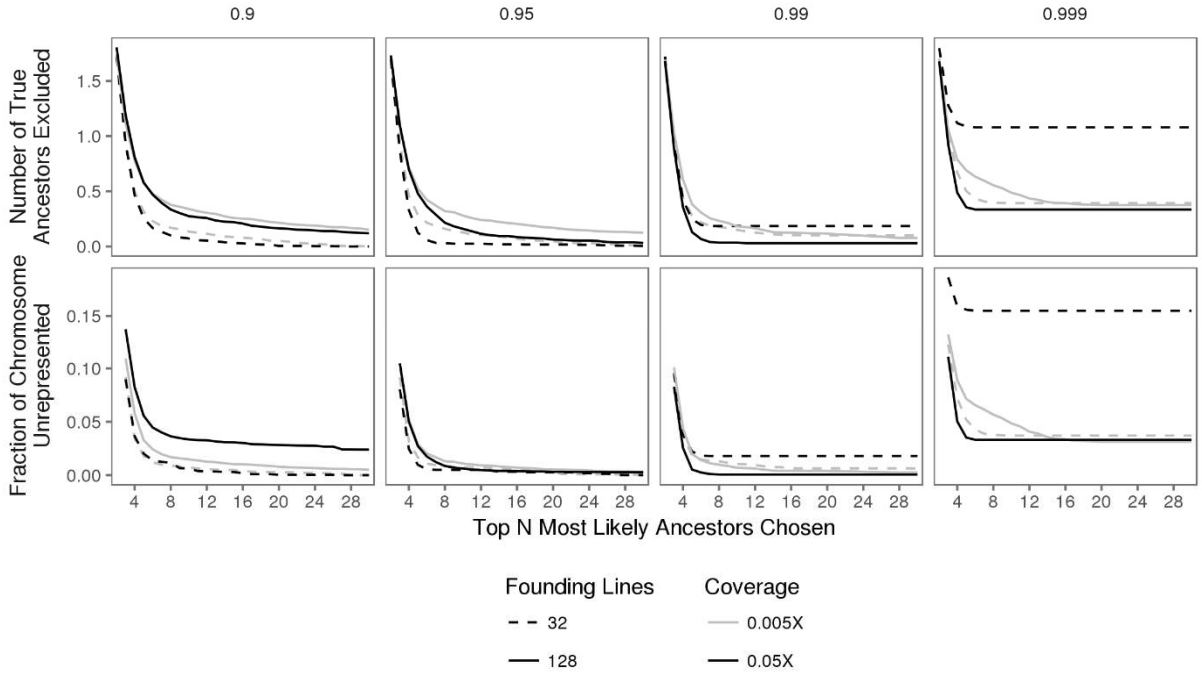

Figure S4. Optimization curves for Most-Likely-Ancestor (MLA) selection for DGRP-derived F_5 hybrid swarm individuals.
Effectiveness is shown for populations founded by either 32 or 128 inbred founding lines across a range of HARP threshold values (0.9 to 0.999), for two levels of sequencing coverage. Values are averaged across 400 reconstructed autosomes (from 100 individuals) per parameter combination.

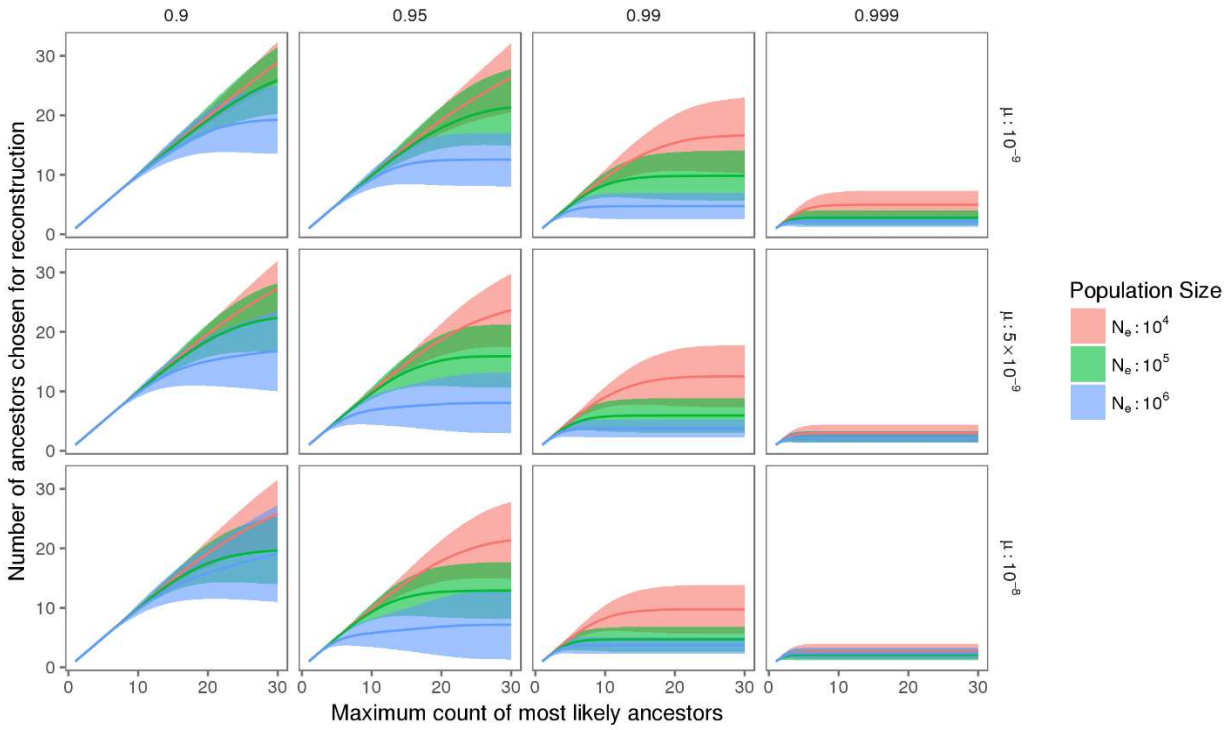

Figure S5. Distribution of Most-Likely-Ancestor counts for simulated, Coalescent-derived, 32-founder F_5 hybrid swarm individuals.
The mean value ± 1 standard deviation is shown by the solid line and ribbon, respectively, across a range of HARP threshold values (0.9 to 0.999), effective population sizes (N_e) and mutation rates (μ). The number of Most-Likely-Ancestors dictates the computational complexity (runtime and memory requirements) of chromosome reconstruction. Each parameter combination includes 400 chromosomes (from 100 simulated individuals) simulated with 0.05X sequencing coverage.

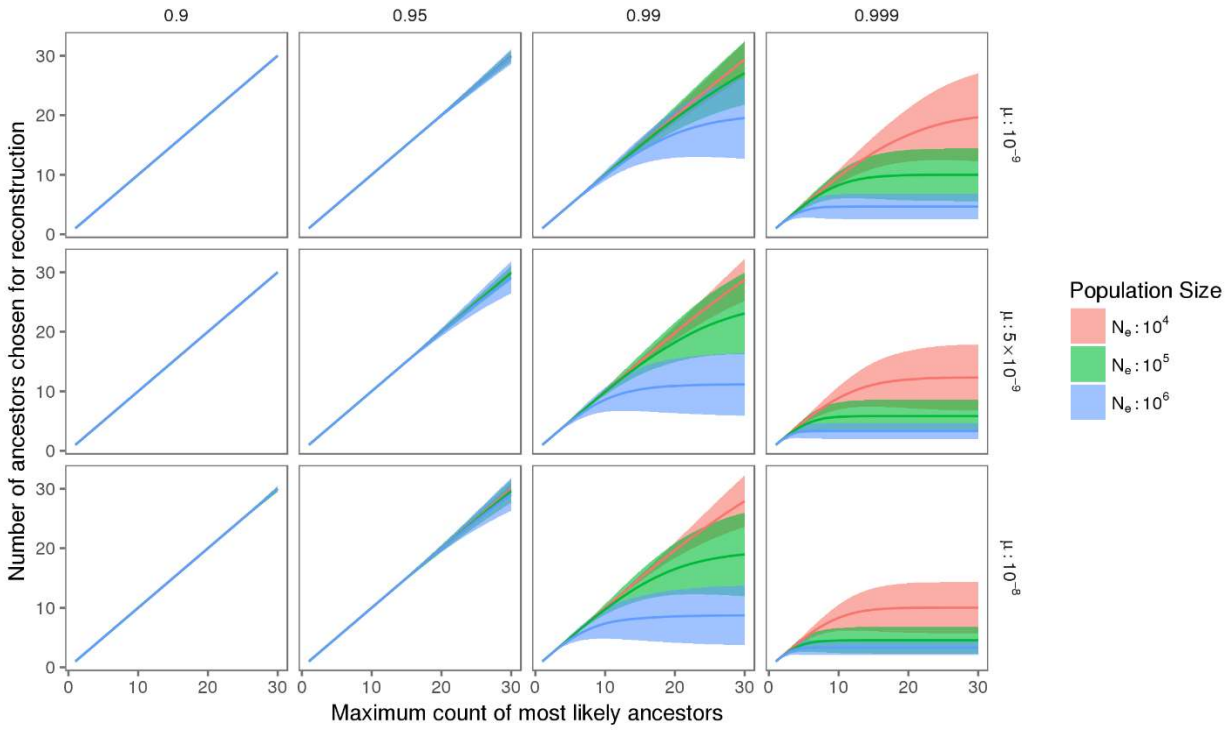

Figure S6. Distribution of Most-Likely-Ancestor counts for simulated, Coalescent-derived, 128-founder F_5 hybrid swarm individuals.
The mean value ± 1 standard deviation is shown by the solid line and ribbon, respectively, across a range of HARP threshold values (0.9 to 0.999), effective population sizes (N_e) and mutation rates (μ). The number of Most-Likely-Ancestors dictates the computational complexity (runtime and memory requirements) of chromosome reconstruction. Each parameter combination includes 400 chromosomes (from 100 simulated individuals) simulated with 0.05X sequencing coverage.

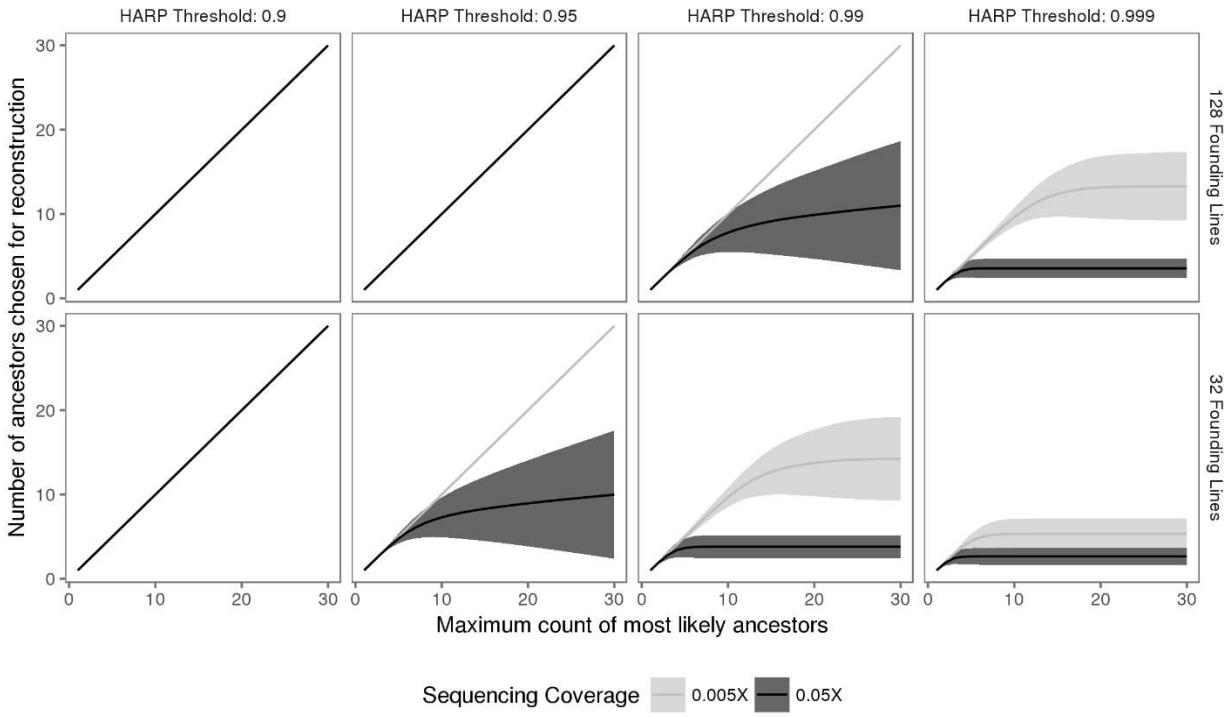

Figure S7. Distribution of Most-Likely-Ancestor counts for simulated, DGRP-derived F_5 hybrid swarm individuals.
The mean value ± 1 standard deviation is shown by the solid line and ribbon, respectively, for populations founded by either 32 or 128 inbred lines, across a range of HARP threshold values (0.9 to 0.999) and two levels of sequencing coverage. Each parameter combination includes 400 chromosomes (from 100 simulated individuals) simulated with 0.05X sequencing coverage.

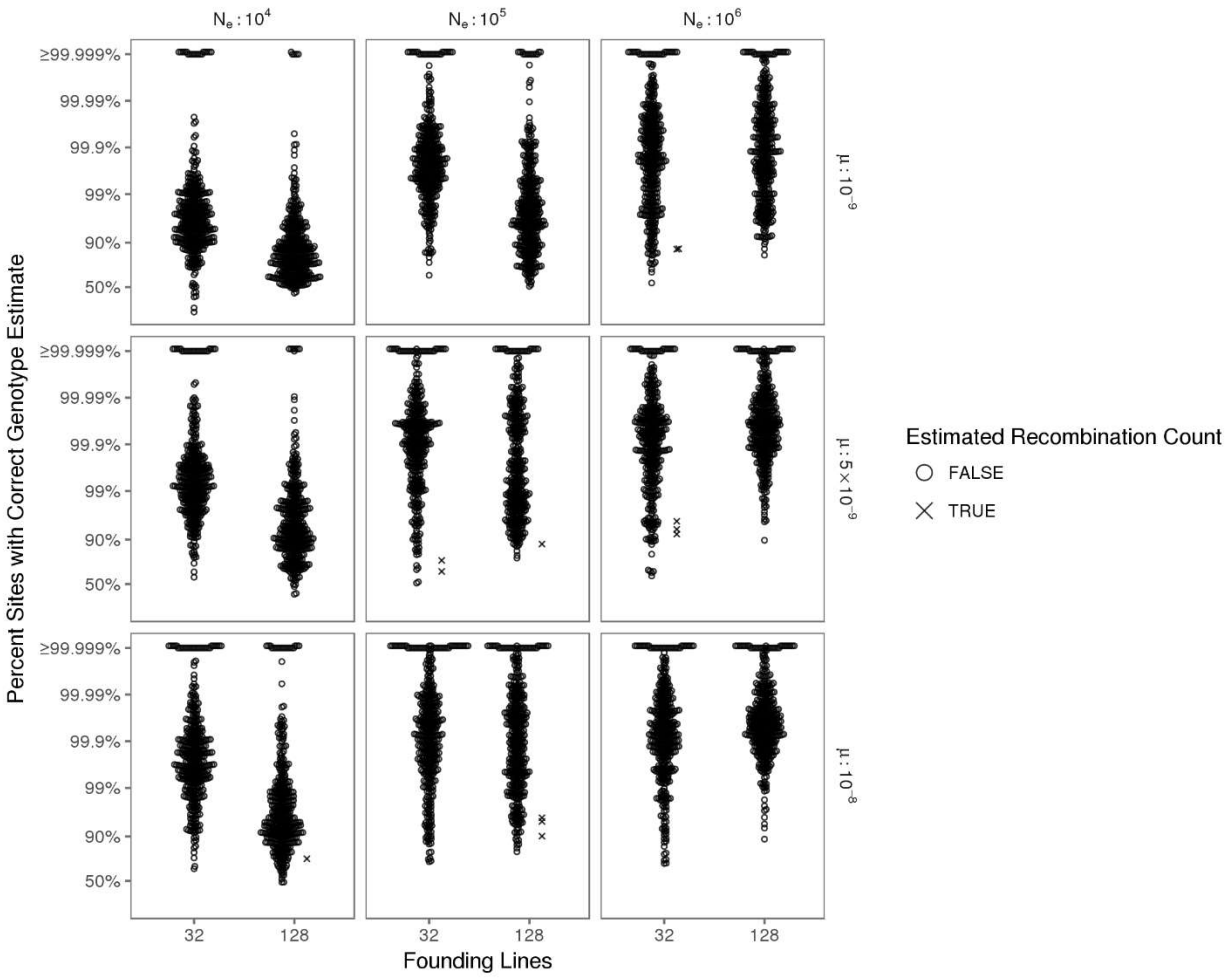

Figure S8. Accuracy of genome reconstruction for simulated, coalescent-derived F_5 hybrid swarm individuals.
Reconstructions were performed for populations simulated as being founded by either 32 or 128 inbred lines for various effective population sizes (N_e) and mutation rates (μ). Accuracy is represented on a logit scale, as most points occur above 90%. Reconstructed chromosomes that are predicted to exhibit ≥10 recombination events are denoted by an X. Each parameter combination includes 400 reconstructed chromosomes (from 100 simulated individuals).

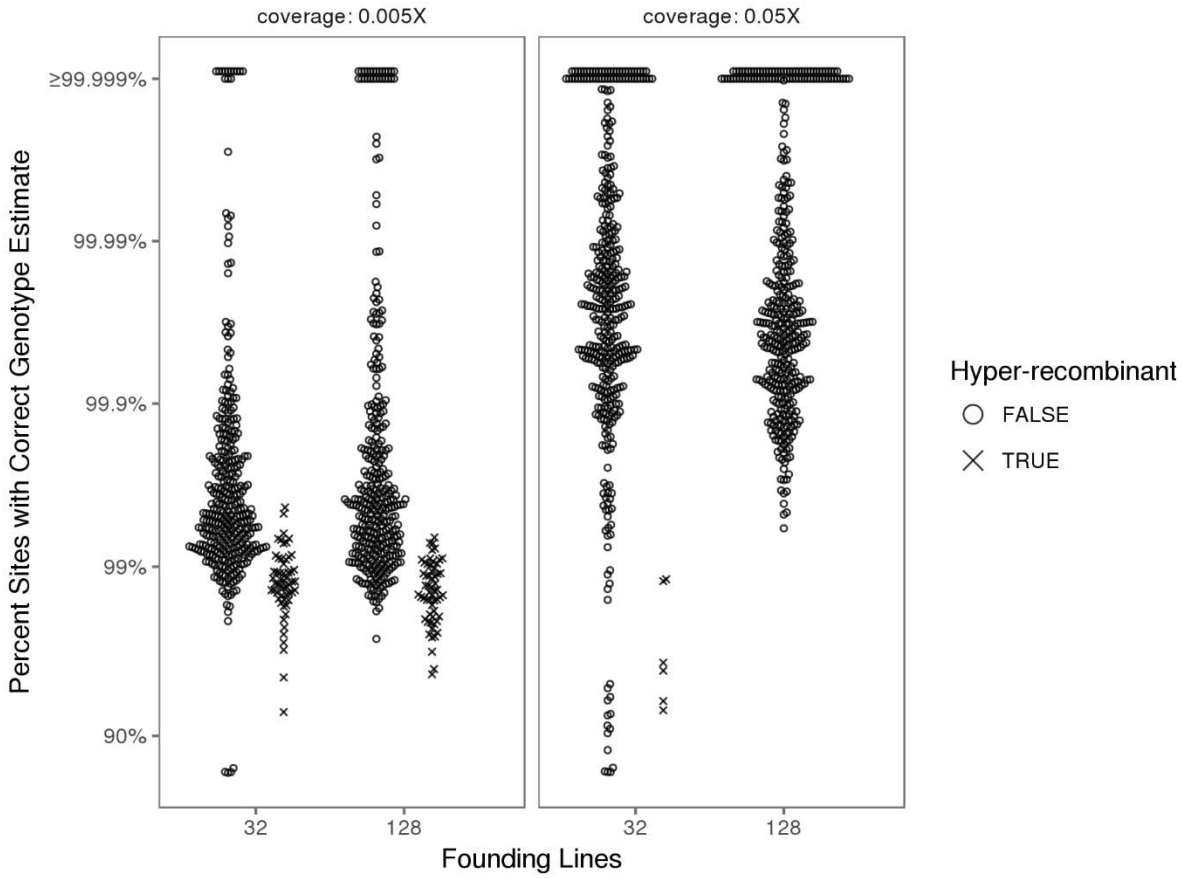

Figure S9 Accuracy of genome reconstruction for simulated, DGRP-derived F_5 hybrid swarm individuals.
Reconstructions were performed for populations simulated as being founded by either 32 or 128 inbred lines for two levels of ultra-low sequencing coverage. Accuracy is represented on a logit scale, as most points occur above 90%. Accuracy values are marked depending on the number of estimated recombination events, with highly recombinant estimates ($\geq$10 recombination events) displayed as an $\times$. Each parameter combination includes 400 reconstructed chromosomes (from 100 simulated individuals).

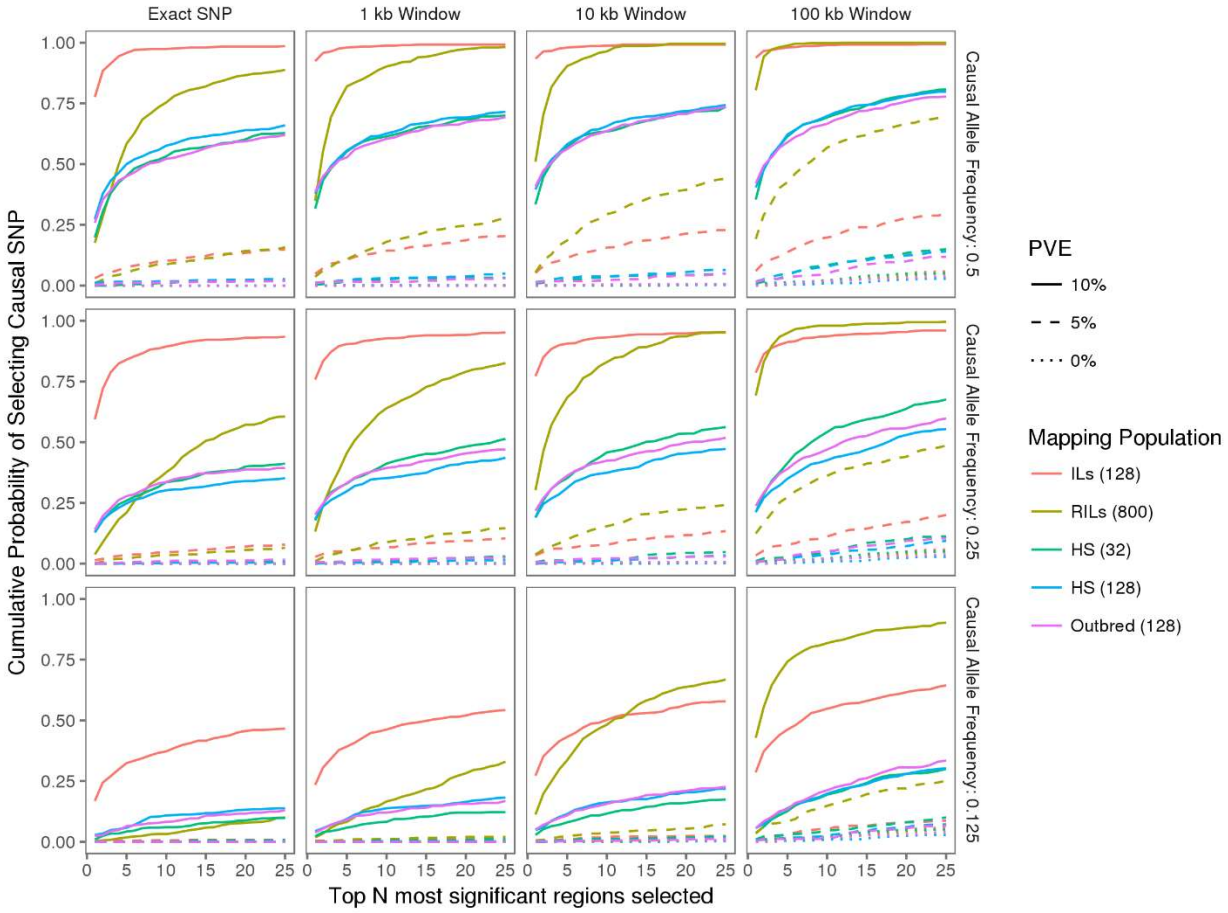

Figure S10. Probability of selecting a causal SNP (or a nearby neighbor) in simulated GWAS.

Each line represents the fraction of GWAS simulations (out of 500 total GWAS, each comprised of 5000 individuals in a case-control framework) where the causal SNP is selected within N most significant regions. Case-control status was assigned based on reference allele dosage at a randomly selected causal SNP segregating with frequency of 50%, 25%, or 12.5%. For 10% PVE, homozygotes for the reference allele were assigned to the case group with 45% probability, while homozygotes for the alternate allele were assigned to the case group with 55% probability (a difference of 10%). For 5% PVE, homozygotes for the reference allele were assigned to the case group with 47.5% probability, while homozygotes for the alternate allele were assigned to the case group with 52.5% probability (a difference of 5%). All heterozygotes (and any individuals modeled with 0% PVE, irrespective of genotype) were equally likely to be assigned to case or control group (a difference of 0%).

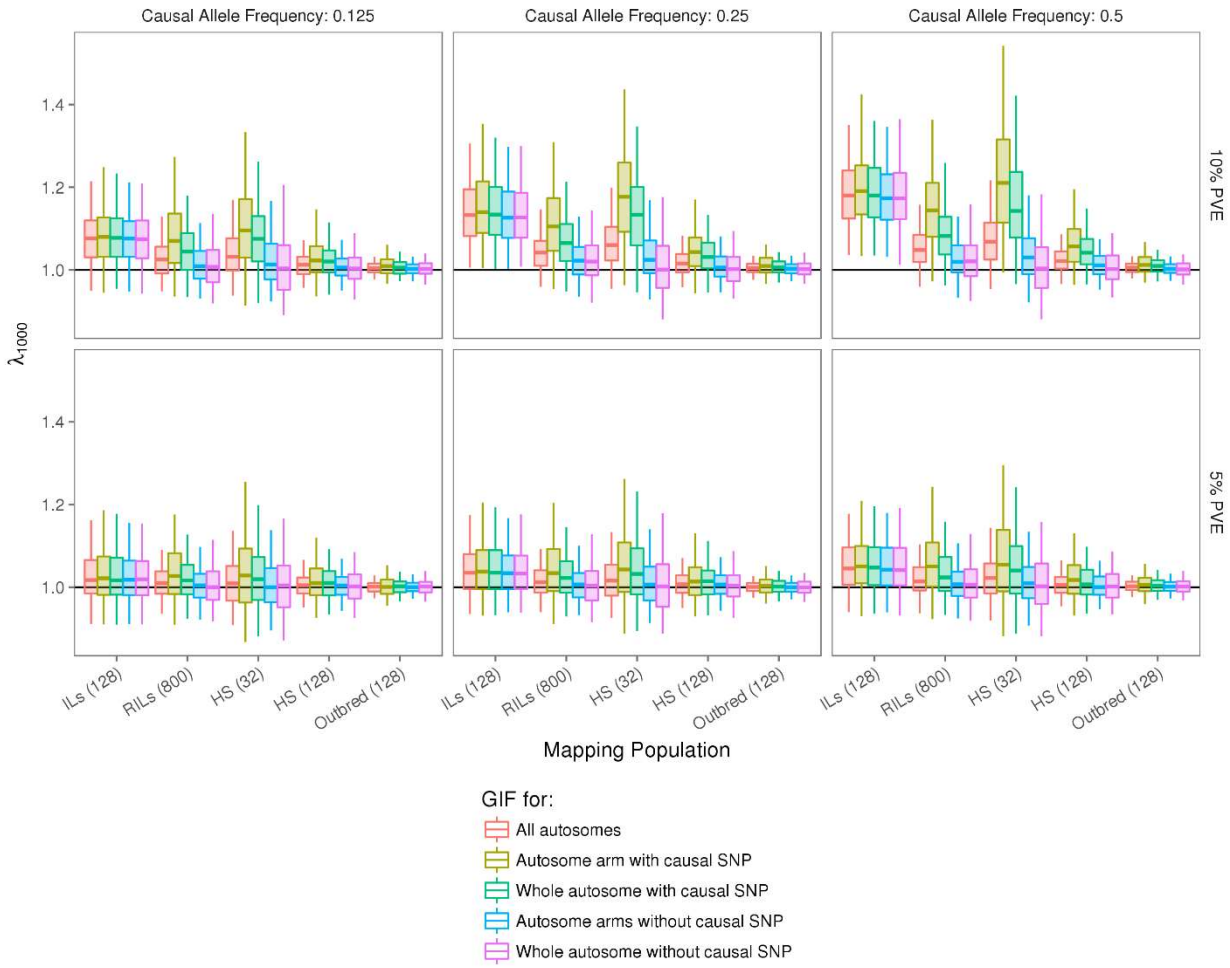

Figure S11. Genomic Inflation Factor (GIF, λ_{1000}) in simulated DGRP GWAS as a function of minor allele frequency and percent variation explained.
GIF is stronger with greater PVE, and is elevated on the arms which contain (and are directly and most strongly linked to) the SNP associated with case and control status.

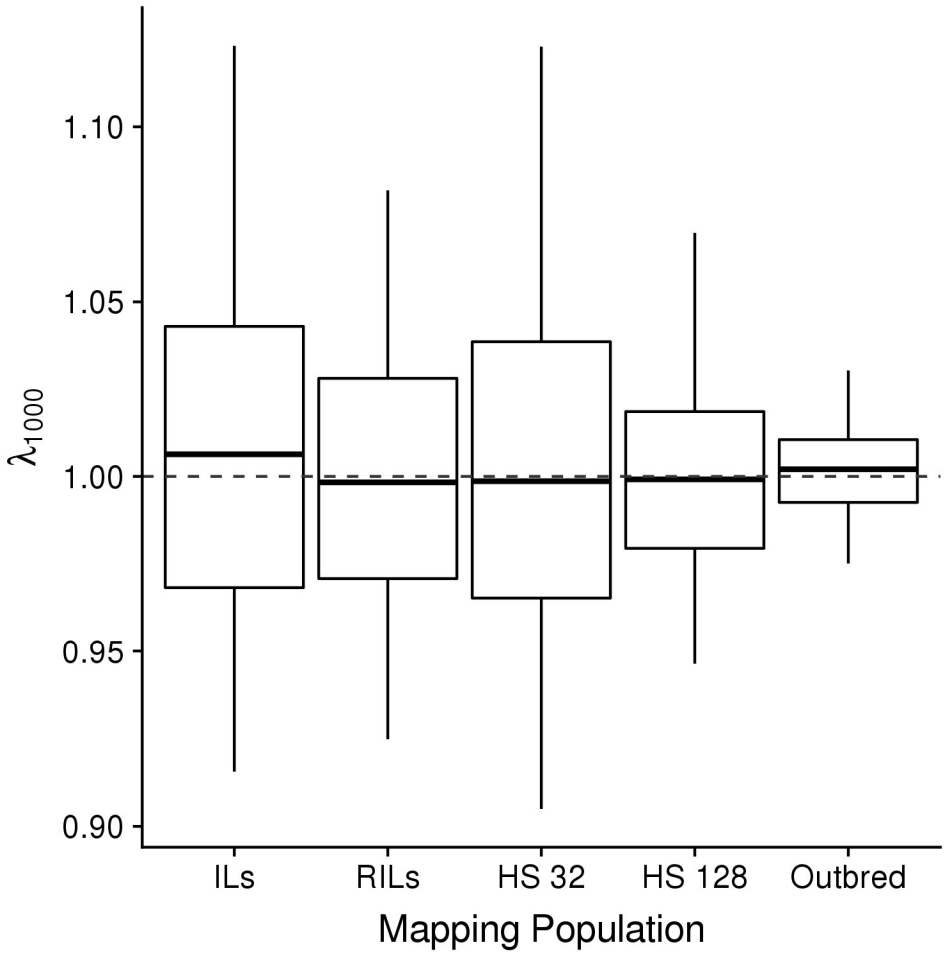

Figure S12. Genomic Inflation Factor (GIF, λ_{1000}) in simulated DGRP GWAS with no genotype-phenotype link.
As the percent variation explained is 0% (and thus, case and control status is randomly assigned), GIF centers around 1.0 as expected.

Population	Coverage	N Founders	N_e	μ	ρ	$\overline{\Delta}$	σ_Δ
DGRP	0.005X	128	-	-	0.201	3.06	3.11
DGRP	0.005X	32	-	-	0.201	3.15	2.81
DGRP	0.05X	128	-	-	0.986	-0.015	0.25
DGRP	0.05X	32	-	-	0.502	0.17	2.15
Coalescent	0.05X	128	10^4	1×10^{-9}	0.029	-2.72	1.60
Coalescent	0.05X	32	10^4	1×10^{-9}	0.219	-1.84	1.50
Coalescent	0.05X	128	10^5	1×10^{-9}	0.288	-1.57	1.64
Coalescent	0.05X	32	10^5	1×10^{-9}	0.761	-0.53	0.99
Coalescent	0.05X	128	10^6	1×10^{-9}	0.742	-0.46	1.09
Coalescent	0.05X	32	10^6	1×10^{-9}	0.754	-0.42	1.22
Coalescent	0.05X	128	10^4	5×10^{-9}	0.203	-2.03	1.73
Coalescent	0.05X	32	10^4	5×10^{-9}	0.622	-0.89	1.17
Coalescent	0.05X	128	10^5	5×10^{-9}	0.652	-0.64	1.41
Coalescent	0.05X	32	10^5	5×10^{-9}	0.782	-0.26	1.16
Coalescent	0.05X	128	10^6	5×10^{-9}	0.956	-0.17	0.44
Coalescent	0.05X	32	10^6	5×10^{-9}	0.759	-0.31	1.26
Coalescent	0.05X	128	10^4	1×10^{-8}	0.238	-1.65	1.86
Coalescent	0.05X	32	10^4	1×10^{-8}	0.745	-0.66	1.05
Coalescent	0.05X	128	10^5	1×10^{-8}	0.776	-0.34	1.12
Coalescent	0.05X	32	10^5	1×10^{-8}	0.846	-0.25	0.93
Coalescent	0.05X	128	10^6	1×10^{-8}	0.937	-0.22	0.56
Coalescent	0.05X	32	10^6	1×10^{-8}	0.833	-0.32	0.95

Table S1. Accuracy of estimated number of recombination events following chromosome reconstruction.
A high concordance correlation coefficient (Lin's ρ) indicates agreement between estimated and true recombination counts for 400 reconstructed chromosomes (coalescent-derived populations) or chromosome arms (DGRP-derived populations). Coalescent-derived populations are described across a range of values for effective population size N_e and mutation rate μ. $\overline{\Delta}$ and σ_Δ denote mean and standard deviation, respectively, for difference between estimated and true recombination count. Reconstructions were performed with a maximum of 16 most-likely-ancestors with a HARP threshold of 0.99 (see methods for more details).