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 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, 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, modeling genetic variation from the Drosophila Genetic Reference Panel as well as variation from neutral simulations encompassing a range of diversity levels. Next, we compared the power and precision of GWAS using the Hybrid Swarm, inbred lines, recombinant inbred lines (RILs), and highly outbred populations across a range of effect sizes and genetic architectures. Our simulations show that these different mapping panels vary in their power and precision, largely depending on the architecture of the trait. Notably, both the Hybrid Swarm and RILs outperform inbred lines for quantitative traits, but not

for monogenic ones. 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, Multiparental 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.

One alternative approach for generating a high-resolution mapping population is to substitute extensive recombination for increased haplotype diversity (Mott *et al.* 2000; Chia *et al.* 2005). 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 (Cao *et al.* 2011; Huang *et al.* 2011; King *et al.* 2012b; MacKay *et al.* 2012; Srivastava *et al.* 2017; Noble *et al.* 2019). However, with the advent of inexpensive DNA sequencing, the cost of individually based genome-wide genotyping becomes defensible, opening the door for a range of experimental designs aimed at addressing open questions in quantitative biology.

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, across a range of genetic architectures, 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 show the feasibility of cost-effective association mapping in a large outbred multi-parental 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 ($\mathrm{N_e} = 10^4, 10^5, 10^6$), mutation rate ($\mu = 10^{-9}, 5 \times 10^{-9}, 10^{-8}$), and number of haplotypes (32, 128). The value θ for each simulation was defined as $4\mathrm{N_e}\mu$. 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 $\times 10^6$ 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 for Reconstruction

Our analyses necessitated a method of storing genotype information for thousands of individuals across thousands of simulations. To do so, we developed a forward-simulator in R that stores ancestral haplotype block maps instead of explicit genotypes. Recombinant individual genotypes can then be reconstituted by extracting ancestral genotypes from ancestor and base pair position indices. This approach leverages information redundancy that exists between related individuals in recombinant populations, achieving between three and four orders of magnitude of compression relative to a VCF file. For example, a compressed VCF file representing a population containing 5000 diploid genotypes 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).

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).

158 For populations founded by simulated haplotypes, recombination occurred in both
159 sexes, with recombination occurring uniformly across each chromosome (Supplemental
160 Figure S1).
161
162 **Simulating and Mapping Sequencing Data**
163 We used *wgsim* (Li 2011) to generate simulated reads. To achieve a desired level of
164 sequencing coverage $C = 0.05$ or 0.005, we generated $N = (C \times S)/(2 \times L)$ reads per
165 chromosome, with read length $L = 100$ bp and chromosome length S bp. We specified a
166 base error rate of 0.001 and an indel fraction of 0. Remaining *wgsim* parameters were
167 left as default.
168
169 We assembled paired end reads using *PEAR* (Zhang *et al.* 2014) and separately
170 aligned the assembled and unassembled groups to a reference genome with *bwa*
171 0.7.14 using the BWA-MEM algorithm (Li 2013). Reads from DGRP-derived populations
172 were mapped to the *D. melanogaster* reference genome v5.39, and reads from
173 coalescent-derived populations were mapped to their respective simulated reference
174 genomes. After converting mapped reads to compressed BAM format with *samtools*
175 1.3.1 (Li *et al.* 2009), we removed PCR duplicates with *Picard tools* 2.0.1 (Broad
176 Institute 2015a).
177
178 **Most Likely Ancestors Selection**
179 To make chromosome reconstructions in the hybrid swarm computationally tractable
180 (Figure 1), we developed a method of accurately selecting a subset of most likely
181 ancestors for any single chromosome. We then used that ancestor subset to reconstruct
182 haplotype blocks using the *RABBIT* package (Zheng *et al.* 2015) in Mathematica.
183 RABBIT operates as a Hidden Markov Model (HMM) using the Viterbi algorithm to
184 return the most likely series of parental combinations (hidden states) across the
185 genome (SNP positions) given the observations (sequenced alleles). For every position
186 in the genome, the Viterbi algorithm evaluates relative likelihoods of transitioning to any
187 possible hidden state. Because the hidden states in our case are ancestor
188 combinations, there will be $(N^2 + N)/2$ combinations of N haplotypes to evaluate at

189 every site. This number of evaluations is tractable at smaller values of N but grows at a
190 quadratic rate. For example, increasing the number of founding haplotypes from 8 to
191 128 is a 16-fold increase in haplotypes, but it would incur orders of magnitude increases
192 in computational effort (Figure 1). Thus, in order to make reconstructions in *RABBIT*
193 computationally tractable for hybrid swarm individuals, it is necessary to identify a
194 subset of founders that accurately includes the true ancestors contributing to any given
195 chromosome.
196
197 We used the software package *HARP* (Kessner *et al.* 2013) to rank the population
198 founding lines based on likelihood of being a true ancestor of a chromosome to be
199 reconstructed. *HARP* was originally developed to estimate haplotype frequencies from
200 pooled sequence data, and we co-opted it to assess relative likelihood that any founder
201 contributed to a genomic window. We ran *HARP* with non-overlapping 100 kb windows
202 with a minimum frequency cutoff 0.0001, producing output which can be visualized as a
203 heat map of ancestor likelihood across the chromosome. A custom *R* script analyzed
204 this *HARP* output and ranked all possible founders in terms of likelihood of contribution
205 for a given chromosome. Briefly, a chromosome-wide significance threshold was
206 calculated, e.g. the 95% or 99% quantile of all likelihoods across all founders and all
207 chromosome windows. Then, every potential ancestor for each 100 kb window was
208 classified as falling above or below this threshold. Founding lines were then ranked in
209 descending order of the number of windows passing the threshold. We examined two
210 measures of effectiveness for this method across a range of quantile threshold values
211 (90%, 95%, 99%, and 99.9%) when selecting up to a maximum number of most likely
212 ancestral haplotypes. The first measure is the number of true ancestral founders
213 excluded; the second measure is the fraction of the chromosome derived from
214 ancestors missing from the selected subset.
215
216 **Chromosome Reconstruction with RABBIT**
217 We used the MAGIC reconstruct method of the Mathematica package RABBIT (Zheng
218 *et al.* 2015) to perform chromosome reconstructions, which has been shown to be
219 accurate for genotype estimation at sequencing coverage at 0.05X for a variety of

220 multiparent populations (Zheng *et al.* 2018). RABBIT requires three inputs: observed
221 genotypes in the individual being reconstructed; map distance (in cM units) of the same
222 loci; and genotypes for the potential ancestors at those same loci. For DGRP-derived
223 simulated populations, we specified map distance based on values reported by
224 Comeron et al. (2012) by performing linear interpolation of cumulative map units (cM) as
225 a function of base pair position. For populations derived from simulated haplotyopes, we
226 used a linear function of 37.5 cM over each 25 Mb chromosome. To specify genotype
227 information, we first counted reference and alternate reads using the Genome Analysis
228 Toolkit *ASEReadCounter* (Broad Institute 2015b). Because it is not possible to make
229 confident homozygote genotype calls from low coverage sequencing data where most
230 sites are observed only once and or twice, we only included diploid genotype
231 observations for sites where both reference and alternate alleles were observed. As
232 RABBIT allows for an ambiguous allele character, for all sites where only reference or
233 alternate reads were observed (but not both), we included one ambiguous allele.
234
235 To minimize memory and runtime requirements, we included at most 5,000 SNPs per
236 chromosome, selected for maximum ancestor-discerning information content. If an
237 observed (sequenced) allele is common, it will only slightly narrow down the possibility
238 of ancestors. If a sequenced allele is rare—at the most extreme, unique to one
239 individual—it provides greater information from which founder that site is derived. Thus,
240 we designate information-rich sites as those where the frequency of the sequenced
241 allele is the lowest with respect to the pool of most likely ancestors. In order to sample
242 sites with high information content spread throughout the chromosome, we used an
243 iterative approach. First, we included all heterozygous sites (i.e. where reference and
244 alternate alleles are both observed). Then, 10% of all SNPs were randomly sampled,
245 and we retained up to the top 0.2% most informative sites, repeating the random
246 sampling and retention until we designated 5,000 SNPs. In principle, this approach
247 could be applied in an iterative manner surrounding recombination breakpoints to fine
248 map their location.
249

250 We ran RABBIT independently for each chromosome using the Viterbi decoding
251 function under the joint model, with all other RABBIT parameters left at default. RABBIT
252 output was converted to a phased chromosome haplotype map, which we then used to
253 extract and concatenate genotype information from a VCF file containing founder
254 genotypes. To calculate genotype reconstruction accuracy, we first imported true
255 (simulated) and estimated (reconstructed) genotypes using a custom *R* script. We
256 measured the fraction of all remaining sites where the estimated diploid genotype is
257 identical to the originally simulated diploid genotype, excluding fixed sites with respect
258 to the founding haplotypes, and excluding any sites with missing genotype information.
259 To ensure our measurement of reconstruction accuracy is consistent across all
260 individuals, we excluded the Y and X chromosomes, which are hemizygous in males.
261
262 To measure accuracy of estimated frequency of recombination events, true and
263 estimated recombination counts were first summed over both copies of each
264 chromosome in a simulated individual. This removed the possibility of introducing error
265 by comparing the wrong copies of chromosomes. Only detectable recombination events
266 were considered, i.e. those that did not occur between homologous haplotypes. We
267 then calculated Lin's concordance correlation coefficient (ρ) between the true and
268 estimated recombination counts using the *epi.ccc* function of the *R* package *epiR*
269 (Stevenson 2018).
270
271 **Chromosome Reconstruction with STITCH**
272 We estimated genotypes with alternative software, STITCH, which is capable of
273 estimating genotypes without reference panels (Davies *et al.* 2016). For assessing
274 STITCH genotype estimate accuracy, we generated and mapped simulated reads for
275 chromosome 2L at 0.05X coverage for 5000 individuals in a F_5 Hybrid Swarm
276 population (N=10000 individuals per generation). Because STITCH draws inference
277 from haplotypes between related individuals, genotype estimate accuracy is expected to
278 increase with greater numbers of sequenced individuals. To capture the effect of
279 sample size on STITCH accuracy, we ran STITCH for sample sizes from 100 to 5000
280 sequenced individuals. We calculated genotype estimate accuracy as the fraction of

281 variable sites with a correct diploid genotype estimate using a custom R script. Because
282 STITCH memory requirements increase with greater values of *k* (founding haplotypes),
283 it was not possible to evaluate *k*=128 founders, or to use the *diploid* estimation model.
284 As a result, we limited our evaluations to 32-founder populations, using the
285 *pseudoHaploid* model.
286
287 **Modeling Computational Complexity of Chromosome Reconstruction**
288 To estimate the rate at which computational requirements grows with data input, we
289 performed chromosome reconstructions with varying numbers of potential founders and
290 markers (SNPs). This allows us to extrapolate the runtime and memory for performing
291 the most resource intensive chromosome reconstructions (i.e. those with > 40 founding
292 lines). To generate runtime and memory usage data, we performed 900 reconstructions
293 using varying sizes of RABBIT input for a single example individual 2L chromosome.
294 Reconstructions included the four true ancestors of the simulated individual, plus 0 to 32
295 additional haplotypes (for a total of between 4 and 36 founders, in steps of 4) and a
296 random selection of marker SNPs (between 500 and 5000 in steps of 500). Ten
297 replicates, each with a unique random set of SNPs, was conducted for each
298 combination of N founding lines and S SNPs using a single core on the University of
299 Virginia computing cluster, with total runtime and peak memory usage as reported from
300 the SLURM workload manager (CPUtime and MaxRSS, respectively). We then
301 modeled the mean runtime and memory usage (averaged across 10 replicates per
302 parameter combination) as a function of number of founding lines and number of SNPs
303 fed into RABBIT. For runtime, simulations involving 8 or fewer founding lines were
304 omitted from the regression model because they ran too quickly to resolve non-zero
305 runtime. Memory was modeled as $Memory(GB) = 7.367 \times 10^{-9} \times SN^4 + 0.0316$, while
306 runtime was modeled as $Runtime(Minutes) = [1.189 \times 10^{-3} \times N^2 + 1.038 \times 10^{-6} \times$
307 $SN^2 + 2.649 \times 10^{-4} \times S]^2$.
308
309 **Simulating Mapping Populations for GWAS**
310 We performed GWAS on mapping populations produced by random sampling and
311 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. For a single simulated mapping population, we began by sampling a random subset of 5,000 individuals, out of 10,000 total individuals generated by forward-simulation. Though our simulations sampled with replacement, this is optional and easily defined during forward-simulation. 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. An overview of our forward-simulator is shown in Supplemental Figure S2.

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; Inbred Lines (ILs) to represent the DGRP; and Recombinant Inbred Lines (RILs), similar to the Drosophila Synthetic Population Resource, or DSPR (King *et al.* 2012a)

The F_{50} populations were generated with 128 founders in same manner as the Hybrid Swarm, except that populations were simulated over 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 DSPR. For computational simplicity, after the 25 generations of isogenization we

343 removed any remaining residual heterozygosity by forcing the identity of a second
344 chromosome copy to be identical to the first copy. We then sampled 5,000 draws (with
345 replacement) of the 800 RILs followed by ancestry permutation as described above.
346
347 To simulate GWAS on Inbred Lines, no forward-simulation was necessary. For a single
348 simulated population, we first randomly selected 128 DGRP lines with high coverage
349 and low levels of heterozygosity as the set of founders. Then, those 128 lines were
350 randomly sampled with replacement 5,000 times. As with hybrid swarm and RILs, for
351 any parameter combination we generated a total of 500 mapping populations.
352
353 **Assigning GWAS Phenotypes**
354 Our GWAS simulation uses a liability model, where an observable binary phenotype is
355 produced from an underlying continuous trait, known as its liability (Xu and Atchley
356 1996). Using the liability framework, we assign case or control phenotypes to mapping
357 populations derived from DGRP chromosome 2L haplotypes. First, a genotype-
358 dependent risk score is calculated for every individual, which is translated into
359 probabilities for case or control group assignment. We parameterized our model such
360 that individuals begin with equivalent odds of being assigned to case or control groups,
361 and that probability is modified depending on allele status at causal loci. For a given
362 simulation, *N* causal loci are randomly selected with equal probability out of all
363 segregating sites in the mapping population. The effect of individual causal loci are
364 simulated as a Gaussian variable drawn with a small extent of noise (sd=0.005) with an
365 expected (mean) value dependent on the allele frequency at that locus. Intermediate
366 frequency loci (i.e. allele frequency = 0.5) are simulated to contribute on average no
367 effect on case/control assignment; alleles near fixation decrease risk; rare alleles
368 increase risk. The linear center of the curve, where most individual risk scores exist
369 under the parameter combinations used in our simulations, approximates additive
370 genetic architecture. The sigmoid tails allow for the continuous risk score to translate to
371 a phenotype necessarily bounded by zero and one. The relationship between allele
372 frequency, risk score, and phenotypic assignment is diagrammed in Supplemental

373 Figure S6. Note that we performed GWAS simulations assuming 100% genotype
374 accuracy.
375
376
377 **Assessing GWAS accuracy**
378 To generate p-values, we performed a χ^2 test of independence for reference and
379 alternate allele counts between case and control groups. To do so most efficiently, we
380 developed a method of aggregating allele counts compatible with our haplotype map
381 format. Briefly, haplotype map breakpoints across all individuals are sorted in ascending
382 order. When iterating through ascending unique start and stop positions, between any
383 pair of breakpoints, all SNPs will be comprised of the same number of each founding
384 haplotype. Haplotype IDs could then be counted and sorted in the same column position
385 order as the table containing polarized allele status (-1 for alternate, +1 for reference).
386 Multiplying the genotype table by the haplotype count vector results in final allele
387 counts, polarized negative for alternate alleles and positive for reference alleles. For
388 inbred mapping populations, we corrected for non-independent allele draws by dividing
389 the χ^2 value by two. Our association testing workflow is visually described in
390 Supplemental Figure S7.
391
392 To measure GWAS accuracy, we generated Receiver Operator Characteristic (ROC)
393 curves and calculated Area Under the Curve (AUC) using the *R* package *pROC* (Turck
394 *et al.* 2011) in the manner described by Marigorta *et al.* (2018). To summarize GWAS
395 accuracy across replicate simulations, we also generated a single representative ROC
396 curve by averaging sensitivity across specificity values.
397
398 **Genomic Inflation Factor**
399 We calculated genomic inflation factor (GIF, λ_{1000}) as the value of $\chi^2_{observed}/\chi^2_{expected}$
400 with two degrees of freedom. Because GIF increases with sample size, we performed a
401 correction to report λ_{1000}, the level of GIF expected with a sample size of 1,000 case
402 and 1,000 control individuals (Freedman *et al.* 2004). Simulations used to calculate GIF
403 were identical to those for simulated GWAS, but included autosome arms 2R, 3L, and

3R in addition to 2L. In order to discriminate site as linked or unlinked to causal loci, all causal loci were forced to exist on the same chromosome.

Data Availability Statement

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

411
412 **Results**
413
414 **Computational Complexity of Chromosome Reconstruction**
415 To determine reasonable limits for numbers of SNPs and haplotypes used for
416 chromosome reconstruction with RABBIT, we modeled peak memory usage and
417 runtime across a range of input sizes. Peak memory grew linearly with number of SNPs
418 used, and at a greater-than-linear rate with haplotypes (Figure 1A). The runtime of
419 RABBIT increased at a greater-than-linear rate for both number of SNPs and number of
420 haplotypes, with the *N* parameter contributing most to resource cost (Figure 1B). These
421 models allowed us to estimate resource requirements at greater numbers of haplotypes
422 (Figure 1, C & D) which would be unfeasible to measure empirically. We find that
423 utilizing more than ~32 founders makes computational costs too demanding, and
424 therefore developed a heuristic rule to identify the most-likely set of ancestors of
425 ancestors.
426
427 **Most-Likely-Ancestor Selection**
428 To reduce computational requirements of haplotype reconstructions with RABBIT, we
429 developed and evaluated an algorithm for selecting a minimum representative set of
430 Most-Likely-Ancestors (MLAs). This step required optimizing a discrimination threshold
431 for values calculated with HARP (Kessner *et al.* 2013, see methods). We found that
432 HARP thresholds of 0.95 and 0.99 discerned a minimal subset of founding lines that
433 include a given chromosome's true ancestors, and represented most of that individual's
434 genomic ancestry (Figure 3). The threshold of 0.99 performed slightly better when 128
435 founding lines were used to found a population, while the threshold of 0.95 performed
436 slightly better with only 32 founding lines. At these thresholds, outcomes were
437 asymptotic by 16 founding lines, a computationally tractable number for RABBIT (Figure
438 1). Populations simulated with genetic variation derived from coalescent models
439 described above included the parameters $N_e = 10^6$ and $\mu = 5 \times 10^{-9}$. HARP thresholds
440 of 0.95 and 0.99 perform similarly well across a range of N_e and μ, suggesting this is a
441 reasonable starting point for optimizing the pipeline for other populations (Supplemental

Figure S3, A). Surprisingly, reducing sequencing coverage by a factor of ten (from 0.05X to 0.005X) resulted in similar levels of ancestor selection accuracy (Supplemental Figure S3, B), demonstrating that additional coverage may not necessarily improve reconstruction accuracy. For simplicity, we next conducted all genome reconstructions with the 0.99 threshold at a simulated sequencing coverage of 0.05X.

Reconstruction Accuracy

Chromosome reconstruction (workflow diagrammed in Supplemental Figure S2) 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, or made use of DGRP or coalescent haplotypes ($\mathrm{N_e} = 10^6$ and $\mu = 5 \times 10^{-9}$). We additionally report reconstruction accuracy in coalescent-derived populations across a range of $\mathrm{N_e}$ and μ values in supplemental Figure S4. As expected, reconstruction accuracy improves with genetic diversity (i.e. greater θ, equal to $4N_e\mu$).

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%.

473 Next, we examined the accuracy of our estimates of recombination number, per
474 individual. The number of recombination events estimated from chromosome
475 reconstruction was most accurate for populations founded by 128 lines (Table 1).
476 Reconstructions of DGRP- and Coalescent-derived chromosomes yielded
477 recombination count estimates that were 98.6% and 95.6% concordant with their
478 respective true recombination counts (Lin's concordance correlation coefficient, ρ).
479 When populations were founded by 32 lines, recombination count estimates were more
480 inaccurate, with DGRP- and Coalescent-derived reconstructions achieving 50.2% and
481 75.9% concordance with their respective true recombination counts. For 32-founder
482 populations, DGRP-derived reconstructions tended to slightly overestimate
483 recombination counts, while the same counts were underestimated for coalescent-
484 derived populations
485
486 Hyper-recombinant estimates (i.e. $\geq$ 10 recombination events for our simulations) were
487 a faithful proxy for inaccurate reconstructions, though inaccurate reconstructions weren't
488 necessarily hyper-recombinant (Figure 4A; Figure S4). The cause of these hyper-
489 recombinant reconstructions is not always clear, but frequently appears due to
490 RABBIT's estimation switching between closely-related haplotypes. Within DGRP-
491 derived 32-founder populations, hyper-recombinant individuals were below the sixth
492 percentile of genotype accuracy (N=6/400 simulations, genotype accuracy
493 range=92.8%-98.8%). For coalescent-derived 32-founder populations, hyper-
494 recombinant individuals fell in the bottom 9% of genotype accuracy (N=3/400
495 simulations, genotype accuracy range = 92.1%-95.6%). For coalescent-derived 32-
496 founder populations, 4.25% (17/400) of reconstructions without hyper-recombinant
497 estimates exhibited lower genotype accuracy than the least accurate hyper-recombinant
498 simulation (range = 59.7%-92.1%). Similarly, for DGRP-derived 32-founder populations,
499 2.5% (10/400) of reconstructions without hyper-recombinant estimates exhibited lower
500 genotype accuracy than the least accurate hyper-recombinant simulation (range =
501 82.3%-92.8%).
502

503 Reducing sequencing coverage by an order of magnitude from 0.05X to 0.005X resulted
504 in more frequent over-estimation of recombination, though overall median genotype
505 accuracy remained above 99% (Figure S5). At 0.005X coverage, hyper-recombinant
506 reconstructions exhibited genotype estimates with accuracy below 99%, while the
507 remaining simulations (with lower recombinant counts) achieved above 99% genotype
508 accuracy. For populations founded by 32 DGRP lines and sequenced at 0.005X
509 coverage, 14% of simulations produced hyper-recombinant estimates (N=56/400), of
510 which only 26.8% (N=15/56) surpassed 99% genotype accuracy (median=98.5%). The
511 remaining 86% of simulations (N=344/400) that were not hyper-recombinant retained
512 greater accuracy, with 89% of simulations resulting in at least 99% genotype accuracy
513 (median=99.5). Increasing the number of founding DGRP lines from 32 to 128 at 0.005X
514 coverage failed to eliminate hyper-recombinant estimates. With 128 founding lines,
515 14.5% of simulations were hyper-recombinant (N=58/400), of which 24.1% (N=14/58)
516 surpassed 99% genotype accuracy (median=98.6%). The 85.5% of simulations that
517 were not hyper-recombinant (N=342/400), exhibited accurate genotype estimates, with
518 86.5% (296/342) of simulations achieving over 99% genotype accuracy (median=99.6).
519
520 We also performed genotype estimation using STITCH (Davies *et al.* 2016, see
521 Materials and Methods for computational limitations). As expected, STITCH accuracy
522 improved with greater sample size of *N* sequenced individuals (Figure 5), providing
523 68.5% genotype accuracy with *N*=100, increasingly approximately linearly until 94.5%
524 accuracy at *N*=3000, and providing 99% accuracy with *N*=4000.
525
526 **GWAS Accuracy**
527 GWAS accuracy was evaluated from the Area Under the Curve (AUC) of Receiver
528 Operator Characteristic (ROC) curves (Figure 6). An AUC of 1.0 represents a 100%
529 accurate classifier where the most significant p-values from GWAS are causal loci,
530 while an AUC of 0.5 would be expected from randomly permuted data. Because the
531 modeled genotype-phenotype strength is determined by a user-defined coefficient for
532 effect size, we are not concerned with the specific AUC values reported here—but
533 rather the relative ranking across population types.

Hybrid Swarm populations resulted in GWAS performance at least as accurate as any other population tested (Figure 6). Across all modeled architectures, 128-founder Hybrid Swarms yielded GWAS accuracy similar to that of the F50 populations. For single-locus traits, 32-founder Hybrid Swarms outperformed 128-founder populations. However, as simulated traits become more complex, involving more loci, 128-founder populations become more accurate.

Recombinant Inbred Lines performed well across the entire spectrum of parameters evaluated, and this advantage was more pronounced for modeled traits of small effect (Figure 6C, E). The only category for which RILs were outperformed was for single-locus traits of large effect (Figure 6B), for which Inbred Lines and 32-founder Hybrid Swarms fared best..

Inbred Lines yielded the most accurate single-locus GWAS, but also the least accurate multi-locus GWAS results (Figure 6). This multi-locus handicap for Inbred Lines was exacerbated when traits were modeled with larger effects. We suspect that the decreased accuracy of association mapping using inbred lines is due to linkage to causal sites, as well as spurious long-distance linkage (see below).

We also evaluated the GWAS accuracy that results when crossing 32 or 128 founders to generate F1 and F2 mapping populations (Supplemental Figure S7). For single-locus simulations, the number of generations of recombination did not influence accuracy, although 32-founder populations performed better than 128-founders. For multi-locus simulations, GWAS accuracy improved with greater generations of recombination, and this effect was more pronounced with 10-locus traits compared to 5-locus traits.

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

565 into groups randomly, i.e. allele state at a causal SNP dictates nonrandom group
566 assignment, then χ^2 values for that SNP should be inflated to some extent. Nonrandom
567 associations between a causal SNP and other loci can inflated test statistics across a
568 chromosome, or across a whole genome. The genome-wide inflation factor (λ) can be
569 expressed as the ratio of observed and expected median χ^2 values, and represents the
570 extent of inflated association signal. Because the median expected χ^2 statistic
571 increases with sample size, we report λ_{1000}, a sample-size-corrected value that is
572 comparable across studies (Freedman *et al.* 2004). We calculated λ_{1000} aggregated
573 across three groups: genome-wide (across all autosomes); on the autosome arm
574 containing the causal allele (linked); and for sites on the autosome physically unlinked
575 to the causal allele.
576
577 The relative ranking of GIF remained the same between a mono- and polygenic
578 architecture with Inbred Lines producing the greatest GIF, and outbred F50 populations
579 producing the lowest GIF (Figure 7). Interestingly, Inbred Lines display elevated GIF
580 genome-wide, even for sites physically unlinked to causal loci, suggesting spurious
581 long-distance linkage disequilibrium. The Hybrid Swarm, propagated for at least two
582 generations, was able to decrease GIF at unlinked loci and by five generations, the
583 problem was ameliorated.
584
585
586 **Discussion**
587 Here, we show that the Hybrid Swarm is a viable method for association mapping using
588 outbred population. We first demonstrate that ultra-low coverage (0.05X), individual-
589 based whole genome sequencing can be used to reconstruct accurate diploid genomes
590 for species across a range of diversity levels (Figure 4, Supplemental Figure S4).
591 Genome reconstruction accurately infers a reduced subset of most-likely-ancestors
592 (Figure 3) and is computationally efficient (Figure 1). Unlike STITCH, reconstruction
593 accuracy for our pipeline is not dependent upon the number of individuals genotyped
594 (Figure 4). Next, we demonstrate that association mapping using the Hybrid Swarm has
595 GWAS performance comparable to other mapping methods (Figure 6), including various

596 types of inbred lines and completely outbred populations. Lastly, we have described a
597 set of computational tools to simulate various mapping populations (including genome
598 reconstructions, and GWAS), enabling this method to easily be applied to a variety of
599 organisms.
600
601 **Considerations of the Hybrid Swarm approach**
602 The Hybrid Swarm approach is applicable to a wide variety of organisms and
603 experimental designs, conferring potential benefits over inbred reference panels. These
604 benefits are realized in three primary ways by: (1) allowing researchers to address
605 questions that require heterozygotes; 2) reducing labor and the influence of block-
606 effects; and 3) breaking down linkage-disequilibrium and population structure that arises
607 from the sampling of individuals from the wild. These benefits are possible due to the
608 ability to accurately reconstruct genomes from ultra-low-coverage sequencing data, i.e.
609 from cost-effective diluted DNA library preparation for a large number of individuals
610 (Baym *et al.* 2015).
611
612 We evaluated simulated Hybrid Swarms founded by 32 or 128 founders to determine
613 the effect of founding haplotype count on reconstruction accuracy and GWAS
614 performance. One might expect 32-founder populations to yield more accurate
615 reconstructions than 128-founder populations due to greater likelihood of selecting a
616 true ancestor from a smaller pool of ancestors. Yet, 32-founder populations were less
617 accurate in our DGRP simulations (Figure 4), and for some coalescent-based
618 simulations (Supplemental Figure S4). The difference in reconstruction accuracy is
619 perhaps due to the inverse relationship between SNP frequency and information
620 content. In a 128-way cross, rare information-rich SNPs are only likely to exist (and be
621 sampled during low-coverage sequencing) if nucleotide diversity is high. The DGRP
622 haplotypes happen to exist in a parameter space with higher nucleotide diversity,
623 resulting in 128-way reconstructions being more accurate. For GWAS, 32-founder
624 populations tended to perform better than 128-founder populations if the simulated trait
625 was controlled by a single locus. Taking these two factors into consideration, broadly
626 speaking, a Hybrid Swarm with fewer founding haplotypes may be more appropriate for

reconstructing genomes when genetic diversity in the founding lines is low, or if the trait of interest is presumably controlled by one or few loci. Conversely, a greater number of founders may be desired when genetic diversity is high, or the trait of interest is likely to be complex.

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. Phasing the outbred founders could be accomplished in two basic ways. For instance, phased genomes could be generated using a variety of long-read sequencing technologies (Pollard *et al.* 2018). Alternatively, trio- and quartet-phasing methods (Patterson *et al.* 2015), coupled with high coverage sequence data from the outbred parents plus a limited (≤2) number of F1 offspring, would enable accurate phasing of parents to be used for reconstruction of downstream recombinant genomes.

One experimental consideration is the effects of drift and selection that operates during the generation of the Hybrid Swarm. However, it is not clear that such processes will affect the ability to reconstruct genomes nor the specific outcome of GWAS, unless the trait of interest was under strong selection itself. What may be more likely is the reduction of haplotypic diversity due to drift or selection will change the power to detect associations, as we see in the differences between 32- and 128-way crosses (Supplemental Figure S7). The distribution of haplotypes can also be skewed by line-specific differences in fitness or fecundity, with such differences being observed for DGRP lines (Horváth and Kalinka 2016). To attenuate haplotype dropout, it may be prudent to seed a Hybrid Swarm with a large population of directed F1 hybrids produced by round-robin crosses. The F1 population would then be followed by a limited number of generations (e.g., 4-5) of random outbreeding. The production of directed F1 hybrids would facilitate the phasing of parents, in the case that they were themselves outbred.

656 *Representation of heterozygotes*
657 One clear difference between inbred and outbred mapping populations is the presence
658 of heterozygotes. On the one hand, the presence of heterozygotes in outbred
659 populations decreases power to detect association relative to inbred lines for an (semi-)
660 additive allele with a given effect size (Figure 6). However, the reduced statistical power
661 of association mapping in outbred populations may be ameliorated by reduced
662 inbreeding depression (Lee *et al.* 2017) and by the ability to assess the heterozygous
663 effects of alleles.
664
665 The ability to assess heterozygous effects of alleles will provide valuable insights into
666 several interesting aspects of biology, such as the nature of dominance and the identity
667 of regulatory polymorphisms governing allele-specific expression in a variety of
668 heterogenous genomes. An increased understanding of dominance relationships and
669 regulatory polymorphisms is important for advancing our understanding of quantitative
670 trait variation and evolution. For instance, several theoretical models have shown that
671 context dependent dominance of quantitative fitness traits can underlie the stable
672 maintenance of polymorphisms subject to seasonally variable (Wittmann *et al.* 2017) or
673 sexually antagonistic (Connallon and Chenoweth 2019) selection. The ability to
674 efficiently map loci with context dependent dominance relationships will aid in the
675 understanding of the stability and abundance of polymorphisms maintained by these
676 forms of balancing selection.
677
678 Regulatory polymorphisms are known to underlie genetic variation in expression (Brem
679 *et al.* 2002; Cavet *et al.* 2003; Rockman and Kruglyak 2006) and this expression
680 variation can potentially be resolved to exact nucleotide differences (Grosveld *et al.*
681 1987; Rave-Harel *et al.* 1997; Bosma *et al.* 2002). The resulting differences in
682 expression can manifest as phenotypic changes to drive local adaptation (Kudaravalli
683 *et al.* 2009; Fraser *et al.* 2010; Fraser 2011, 2013). Allele-specific expression (ASE)
684 arising from cis-acting regulatory factors is a common mechanism to produce heritable
685 differences in expression (Yan *et al.* 2002; Cowles *et al.* 2002; Lo *et al.* 2003; Doss
686 2005). Because allelic expression biases are only produced (and detectable) in

heterozygotes, Hybrid Swarm populations facilitate the study of regulatory genetic variation (i.e. ASE) as a driver of local adaptation in a variety of organisms.

Undirected outbreeding in a common environment

The Hybrid Swarm approach involves propagation of a single large outbred population, primarily via undirected crossing. This design confers benefits over alternatives of either rearing inbred lines separately or performing controlled crosses. The relative value of these benefits may vary across organisms and experimental designs.

Rearing organisms in a common environment reduces the influence of random block effects that result from rearing in separate enclosures or defined areas. Therefore, studies examining genotype-by-environment interactions can dramatically expand the number of environmental factors used because all individuals were previously reared in a common environment, enabling a Hybrid Swarm to be randomly scatted across environmental treatments. This feature is in contrast to rearing each inbred family across all environments, a practice which generally limits the number of environmental treatments used. Rearing individuals in a common environment also likely reduces genotype-environment correlation that could be exacerbated by vial effects.

Random outbreeding of a single population can require less labor compared to performing controlled crosses or serial propagation of inbred lines. The relative value of these benefits is likely to be species specific. For instance, for larger animals (e.g., mice), controlled crosses may be logistically easier than a randomly mating swarm. For many other species, however, the Hybrid Swarm would be logistically advantageous. The cost of maintaining inbred reference panels also varies among taxa, most notably between those species which can be kept as seeds or in a cryogenically preserved state. For other species, such as flies, the monetary and environmental cost of maintaining hundreds to thousands of inbred-lines may be prohibitive. The maintenance of large panels of inbred lines also poses the risk of contamination, necessitating frequent re-genotyping to confirm fidelity to previously sequenced genomes.

One potential drawback of the randomly outbred method is the aforementioned loss of haplotypes due to genetic drift. If such dropout is a major concern, it could be attenuated by directed F1 mating, followed by random crosses for the subsequent generations. A second drawback of the Hybrid Swarm is the lack of genetic replicates for phenotyping. Inbred panels allow for a single genotype to be phenotyped multiple times, reducing the effects of error associated with phenotyping and micro-environmental variation (Mackay and Huang 2018). Because this benefit is typically not possible in the Hybrid Swarm (unless clones can be propagated), the Hybrid Swarm method represents a tradeoff, reducing the influence of block effects while increasing error associated with phenotyping.

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. Therefore, the Hybrid Swarm approach allows researchers to tailor the mapping population to the traits of interest, by perhaps selecting founders from diverse origins to test specific hypotheses about the distribution of functional variation in the wild.

A limited number of rounds of recombination during the propagation of the Hybrid Swarm also reduces long-distance linkage disequilibrium within a panel of inbred lines derived from a single locality. Long-distance LD results from correlated occurrence of rare variants (Huang *et al.* 2014), potentially contributing to false positives in GWAS (Figure 6). This is reflected in our simulations by genome-wide inflation of λ, even

749 across physically unlinked chromosomes, whereas two or more generations of
750 recombination were sufficient to reduce this inflation (Figure 7).
751
752 **Computational Considerations**
753 The simulations conducted for this analysis were made feasible by three primary
754 innovations. First, the haplotype block file format allowed us to leverage information
755 redundancy between related individuals and store highly compressed, lossless
756 genotype information. With nearly 1/2000th the file size of a compressed VCF file,
757 haplotype block files greatly reduced both the disk storage footprint and time required
758 for disk write operations. Second, instead of performing forward-time simulations for
759 every single iteration of GWAS, permuted subsets of simulated populations allowed for
760 more rapid GWAS simulations. The format of haplotype block files facilitated
761 permutations of the ancestry contained within a population's mosaic haplotypes,
762 generating novel population genetic structure while preserving the forward-simulator's
763 influence of drift and meiotic recombination (Figure 2). Third, instead of extracting site-
764 specific genotypes for every individual in the GWAS analysis, we decreased the number
765 of computational operations by performing aggregate counts across all sites between
766 adjacent recombination events in the population (Figure 3).
767
768 Imputation is commonly combined with low or reduced-representation sequencing as a
769 method of cost-effective genotyping. Software such as Beagle (Browning and Browning
770 2016) and IMPUTE2 (Howie *et al.* 2011) are able to utilize hundreds of representative
771 haplotypes as priors (Howie *et al.* 2012; Das *et al.* 2018), and have only been evaluated
772 with 20 to 200-fold greater sequencing depth, i.e. 1-10X. We chose to compare our
773 reconstruction method to that of STITCH, which was most directly able to interface with
774 our pipeline's BAM files produced from ultra-low-coverage sequencing data. While
775 STITCH was able to accurately estimate genotypes, this was only the case for large
776 sample sizes, while our pipeline's genotype estimation accuracy is independent of
777 sample size (Figure 4).
778

779 The process of haplotype reconstruction in outbred individuals as mosaics of ancestral
780 haplotypes is commonly limited to no more than a dozen founder haplotypes, because
781 every diplotype combination is evaluated (Gatti *et al.* 2014; Zan *et al.* 2019).. For our
782 reconstruction method to be computationally tractable, the subset of most-likely-
783 ancestors must be small (less than 20 or so) and include accurately-inferred ancestors.
784 Should the number of recombination events grow too numerous, this condition becomes
785 more difficult to enforce. This fact underscores the importance of choosing an
786 appropriate number of generations for crossing, informed by known or estimated rates
787 of recombination.
788
789 To simplify computational requirements during GWAS, we chose to simulate
790 case/control phenotypes. This allowed for rapid association tests of our simulated
791 GWAS (Figure 5), enabling us to conduct power analyses across a range of conditions.
792 Because these discrete phenotypes were the product of an underlying quantitative trait
793 (Supplemental Figure S6), we effectively modeled an additive trait with low penetrance.
794 Our observed patterns of GWAS accuracy (Figure 6) should therefore apply to
795 quantitative traits in general. Further, our GWAS simulations were conducted with exact
796 genotypes, which would be analogous to presuming genotype calls for the mapping
797 population were 100% accurate. While our method — or any other method — cannot
798 estimate genotypes with such precision, our observed genotype error rate (0.01% on
799 average, Figure 4), distributed randomly across all individuals in a mapping population,
800 should not influence the conclusions of our observed GWAS accuracy in any
801 meaningful way.
802
803 The use of haplotype-based association tests, instead of the SNP-based statistics we
804 employed, can achieve increased power (albeit with reduced mapping resolution), by
805 performing tests on haplotypes across genomic windows (Long *et al.* 2014). While the
806 utility of haplotype-based methods in the Hybrid Swarm is not known with certainty, we
807 speculate that such haplotype-based methods would be less effective for the Hybrid
808 Swarm compared to inbred populations comprised of a limited number of unique
809 haplotypes. The utility of haplotype-based tests will depend on the number of unique

haplotypes founding a mapping population; as more haplotypes are included, the population will more closely resemble a population composed of (nearly) unique haplotypes. In the parameter regime described here (thousands of individuals sampled from 32- or 128-way crosses), any haplotype would be expected to segregate at < 0.05% (and unique diplotypes even more rare), yielding low statistical power for haplotype-based association testing.

Simulating the Hybrid Swarm Approach

At minimum, the Hybrid Swarm approach requires a sequenced set of individuals for founding a recombinant population. Although our simulations presented here were conducted with inbred founding lines, genome reconstructions can similarly be performed with any phased genomes. For example, 16 phased outbred founders could be treated as 32 independent haplotypes. Optionally, a recombination rate map for the population can be provided, otherwise recombination is assumed to occur with equal user-defined probability across any chromosome.

As a first step, power analyses using our rapid association test simulation pipeline will inform choices of sample size and mapping population design (Figure 3). After determining a feasible sample size for a given SNP of minimum percent variation explained, researchers can evaluate the accuracy of low-coverage chromosome reconstructions for a simulated proposed mapping population. Note that while we 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 (Li *et al.* 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,

841 reconstruction accuracy was low for coalescent-derived mapping populations modeled
842 with $\theta = 4 \times 10^{-5}$ (Supplemental Figure S8), which may reflect those of *C. elegans*
843 (Barriere and Félix 2005). Further, 5000 SNPs may be an over- or under-estimate of
844 those required in other systems. Because recombination between haplotypes can only
845 be inferred at sampled variable sites, SNP density directly influences how close inferred
846 breakpoints will be resolved with respect to their actual position. However, our choice of
847 5000 SNPs is a conservative threshold; RABBIT can accommodate more, depending on
848 SNP density and recombination rate, with modest (linear) increases in computational
849 resources. The models described in Figure 1 can be used to estimate the memory and
850 runtime required for a given number of input ancestor haplotypes. Note that while our
851 pipeline is able to accommodate arbitrarily large numbers of founding haplotypes, this is
852 only because a small number of generations ensures that any single chromosome is
853 composed of a small subset of those founders. As the number of generations increases,
854 so too will the number of ancestors contributing to recombinant chromosomes. If
855 recombination rates are known, one can model recombination events as a Poisson
856 process depending on the number of generations of recombination. From the resulting
857 model, a ceiling of number of generations can be chosen that ensures with some
858 probability that any single chromosome is a combination of ≤ 16 (or any other threshold)
859 ancestors.
860
861 To evaluate whether low coverage sequencing data will yield accurate genotype
862 estimates for a given proposed mapping population, researchers can test reconstruction
863 accuracy *in silico*. We provide a convenient forward-simulation R script for this purpose
864 that generates output in the haplotype map format (Figure 2). Simulated individuals can
865 then be ran through the simulated sequencing and mapping pipeline at a desired level
866 of coverage. After generating simulated mapped individuals, researchers can optimize
867 the number of MLAs and HARP threshold that provide most effective MLA selection for
868 their mapping population (Figure 4). This step may reveal haplotypes that are
869 consistently problematic or inaccurately chosen, which can be excluded from further
870 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 coverage 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.

894 **References**

895 Barriere A., and A.-M. Félix, 2005 Natural variation and population genetics of
896 Caenorhabditis elegans. WormBook.

897 Baym M., S. Kryazhimskiy, T. D. Lieberman, H. Chung, M. M. Desai, *et al.*, 2015
898 Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS
899 One 10: 1–15.

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

903 Bosma P. J., J. R. Chowdhury, C. Bakker, S. Gantla, A. de Boer, *et al.*, 2002 The
904 Genetic Basis of the Reduced Expression of Bilirubin UDP-Glucuronosyltransferase
905 1 in Gilbert's Syndrome. N. Engl. J. Med. 333: 1171–1175.

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

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

909 Broad Institute, 2015b Genome Analysis Toolkit: Variant Discovery in High-Throughput
910 Sequencing Data. https://gatk.broadinstitute.org/hc/en-us

911 Browning B. L., and S. R. Browning, 2016 Genotype Imputation with Millions of
912 Reference Samples. Am. J. Hum. Genet. 98: 116–126.

913 Cao J., K. Schneeberger, S. Ossowski, T. Günther, S. Bender, *et al.*, 2011 Whole-
914 genome sequencing of multiple Arabidopsis thaliana populations. Nat. Genet. 43:
915 956–965.

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

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

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

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

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

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

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

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

934 Das S., G. R. Abecasis, and B. L. Browning, 2018 Genotype Imputation from Large
935 Reference Panels. Annu. Rev. Genomics Hum. Genet. 19: 73–96.

936 Davies R. W., J. Flint, S. Myers, and R. Mott, 2016 Rapid genotype imputation from
937 sequence without reference panels. Nat. Genet. 48: 965–969.

938 Doss S., 2005 Cis-acting expression quantitative trait loci in mice. Genome Res. 15:
939 681–691.

940 Fraser H. B., A. M. Moses, and E. E. Schadt, 2010 Evidence for widespread adaptive
941 evolution of gene expression in budding yeast. Proc. Natl. Acad. Sci. 107: 2977–
942 2982.

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

945 Fraser H. B., 2013 Gene expression drives local adaptation in humans. Genome Res.
946 23: 1089–1096.

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

950 Gatti D. M., K. L. Svenson, A. Shabalin, L.-Y. Wu, W. Valdar, *et al.*, 2014 Quantitative
951 Trait Locus Mapping Methods for Diversity Outbred Mice. G3 4: 1623–1633.

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

955 Horváth B., and A. T. Kalinka, 2016 Effects of larval crowding on quantitative variation
956 for development time and viability in Drosophila melanogaster. Ecol. Evol. 6: 8460–
957 8473.

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

960 Howie B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis, 2012 Fast
961 and accurate genotype imputation in genome-wide association studies through pre-

phasing. Nat. Genet. 44: 955–959.

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

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

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

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

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

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

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

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

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

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

Lee Y. C. G., Q. Yang, W. Chi, S. A. Turkson, W. A. Du, *et al.*, 2017 Genetic architecture of natural variation underlying adult foraging behavior that is essential for survival of Drosophila melanogaster. Genome Biol. Evol. 9: 1357–1369.

Li R., M. A. Lyons, H. Wittenburg, B. Paigen, and G. A. Churchill, 2005 Combining data

998 from multiple inbred line crosses improves the power and resolution of quantitative
999 trait loci mapping. Genetics 169: 1699–709.

1000 Li H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, *et al.*, 2009 The Sequence
1001 Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

1002 Li H., 2011 wgsim (short read simulator)

1003 Li D., J. P. Lewinger, W. J. Gauderman, C. E. Murcray, and D. Conti, 2011 Using
1004 extreme phenotype sampling to identify the rare causal variants of quantitative
1005 traits in association studies. Genet. Epidemiol. 35: 790–9.

1006 Li H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-
1007 MEM

1008 Lo H. S., Z. Wang, Y. Hu, H. H. Yang, S. Gere, *et al.*, 2003 Allelic variation in gene
1009 expression is common in the human genome. Genome Res. 13: 1855–62.

1010 Long A. D., S. J. Macdonald, and E. G. King, 2014 Dissecting complex traits using the
1011 Drosophila Synthetic Population Resource. Trends Genet. 30: 488–495.

1012 Mackay T. F. C., and W. Huang, 2018 Charting the genotype–phenotype map: lessons
1013 from the Drosophila melanogaster Genetic Reference Panel. Wiley Interdiscip. Rev.
1014 Dev. Biol. 7: e289.

1015 MacKay T. F. C. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, *et al.*, 2012
1016 The Drosophila melanogaster Genetic Reference Panel. Nature 482: 173–178.

1017 Marigorta U. M., J. A. Rodríguez, G. Gibson, and A. Navarro, 2018 Replicability and
1018 Prediction: Lessons and Challenges from GWAS. Trends Genet. 34: 504–517.

1019 Monir M. M., and J. Zhu, 2017 Comparing GWAS Results of Complex Traits Using Full
1020 Genetic Model and Additive Models for Revealing Genetic Architecture. Sci. Rep. 7:
1021 38600.

1022 Mott R., C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint, 2000 A method for fine
1023 mapping quantitative trait loci in outbred animal stocks. Proc. Natl. Acad. Sci. 97:
1024 12649–54.

1025 Nikpay M., A. Goel, H. H. Won, L. M. Hall, C. Willenborg, *et al.*, 2015 A comprehensive
1026 1000 Genomes-based genome-wide association meta-analysis of coronary artery
1027 disease. Nat. Genet. 47: 1121–1130.

1028 Noble L. M., M. V Rockman, and H. Teotónio, 2019 Gene-level quantitative trait
1029 mapping in an expanded C . elegans multiparent experimental evolution panel. 1–
1030 7.

1031 Patterson M., T. Marschall, N. Pisanti, L. van Iersel, L. Stougie, *et al.*, 2015 WhatsHap :

1032 Weighted Haplotype Assembly for Future-Generation Sequencing Reads. J.
1033 Comput. Biol. 22: 498–509.

1034 Paul R. Staab, Sha Zhu, Dirk Metzler, and Gerton Lunter, 2015 {scrm}: efficiently
1035 simulating long sequences using the approximated coalescent with recombination.
1036 Bioinformatics 31: 1680–1682.

1037 Pollard M. O., D. Gurdasani, A. J. Mentzer, T. Porter, and M. S. Sandhu, 2018 Long
1038 reads: their purpose and place. Hum. Mol. Genet. 27: R234–R241.

1039 Price A. L., N. A. Zaitlen, D. Reich, and N. Patterson, 2010 New approaches to
1040 population stratification in genome-wide association studies. Nat. Rev. Genet. 11:
1041 459–463.

1042 R Core Team, 2016 R: A Language and Environment for Statistical Computing

1043 Rave-Harel N., E. Kerem, M. Nissim-Rafinia, I. Madjar, R. Goshen, *et al.*, 1997 The
1044 molecular basis of partial penetrance of splicing mutations in cystic fibrosis. Am. J.
1045 Hum. Genet. 60: 87–94.

1046 Rockman M. V., and L. Kruglyak, 2006 Genetics of global gene expression. Nat. Rev.
1047 Genet. 7: 862–872.

1048 Singh R., I. T. Lobina, M. Thomson, S. McCouch, C. Dilla-Ermita, *et al.*, 2013 Multi-
1049 parent advanced generation inter-cross (MAGIC) populations in rice: progress and
1050 potential for genetics research and breeding. Rice 6: 11.

1051 Spencer C. C. A., Z. Su, P. Donnelly, and J. Marchini, 2009 Designing genome-wide
1052 association studies: Sample size, power, imputation, and the choice of genotyping
1053 chip. PLoS Genet. 5.

1054 Srivastava A., A. P. Morgan, M. L. Najarian, V. K. Sarsani, J. S. Sigmon, *et al.*, 2017
1055 Genomes of the Mouse Collaborative Cross. Genetics 206: 537–556.

1056 Stevenson M., 2018 epiR: Tools for the Analysis of Epidemiological Data

1057 Turck N., L. Vutskits, P. Sanchez-Pena, X. Robin, A. Hainard, *et al.*, 2011 pROC: an
1058 open-source package for R and S+ to analyze and compare ROC curves. BMC
1059 Bioinformatics 8: 12–77.

1060 Wittmann M. J., A. O. Bergland, M. W. Feldman, P. S. Schmidt, and D. A. Petrov, 2017
1061 Seasonally fluctuating selection can maintain polymorphism at many loci via
1062 segregation lift. Proc. Natl. Acad. Sci. 114: E9932–E9941.

1063 Wu Y., Z. Zheng, P. M. Visscher, and J. Yang, 2017 Quantifying the mapping precision
1064 of genome-wide association studies using whole-genome sequencing data.
1065 Genome Biol. 18: 1–10.

1066 Xu S., and W. R. Atchley, 1996 Mapping quantitative trait loci for complex binary
1067 diseases using line crosses. Genetics 143: 1417–1424.

1068 Yan H., W. Yuan, V. E. Velculescu, B. Vogelstein, and K. W. Kinzler, 2002 Allelic
1069 variation in human gene expression. Science 297: 1143.

1070 Yang J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price, 2014 Advantages
1071 and pitfalls in the application of mixed-model association methods. Nat. Genet. 46:
1072 100–106.

1073 Yang C., Y. Wang, W. Xu, Z. Liu, S. Zhou, *et al.*, 2018 Genome-wide association study
1074 using diversity outcross mice identified candidate genes of pancreatic cancer.
1075 Genomics 0–1.

1076 Yu J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, *et al.*, 2006 A unified mixed-
1077 model method for association mapping that accounts for multiple levels of
1078 relatedness. Nat. Genet. 38: 203–208.

1079 Zan Y., T. Payen, M. Lillie, C. F. Honaker, P. B. Siegel, *et al.*, 2019 Genotyping by low-
1080 coverage whole-genome sequencing in intercross pedigrees from outbred
1081 founders: A cost-efficient approach. Genet. Sel. Evol. 51: 1–11.

1082 Zhang J., K. Kobert, T. Flouri, and A. Stamatakis, 2014 PEAR: A fast and accurate
1083 Illumina Paired-End reAd mergeR. Bioinformatics 30: 614–620.

1084 Zheng C., M. P. Boer, and F. A. van Eeuwijk, 2015 Reconstruction of genome ancestry
1085 blocks in multiparental populations. Genetics 200: 1073–1087.

1086 Zheng C., M. P. Boer, and F. A. van Eeuwijk, 2018 Accurate genotype imputation in
1087 multiparental populations from low-coverage sequence. Genetics 210: 71–82.

1088

1089

1090 **Figures and Figure Legends**

1091

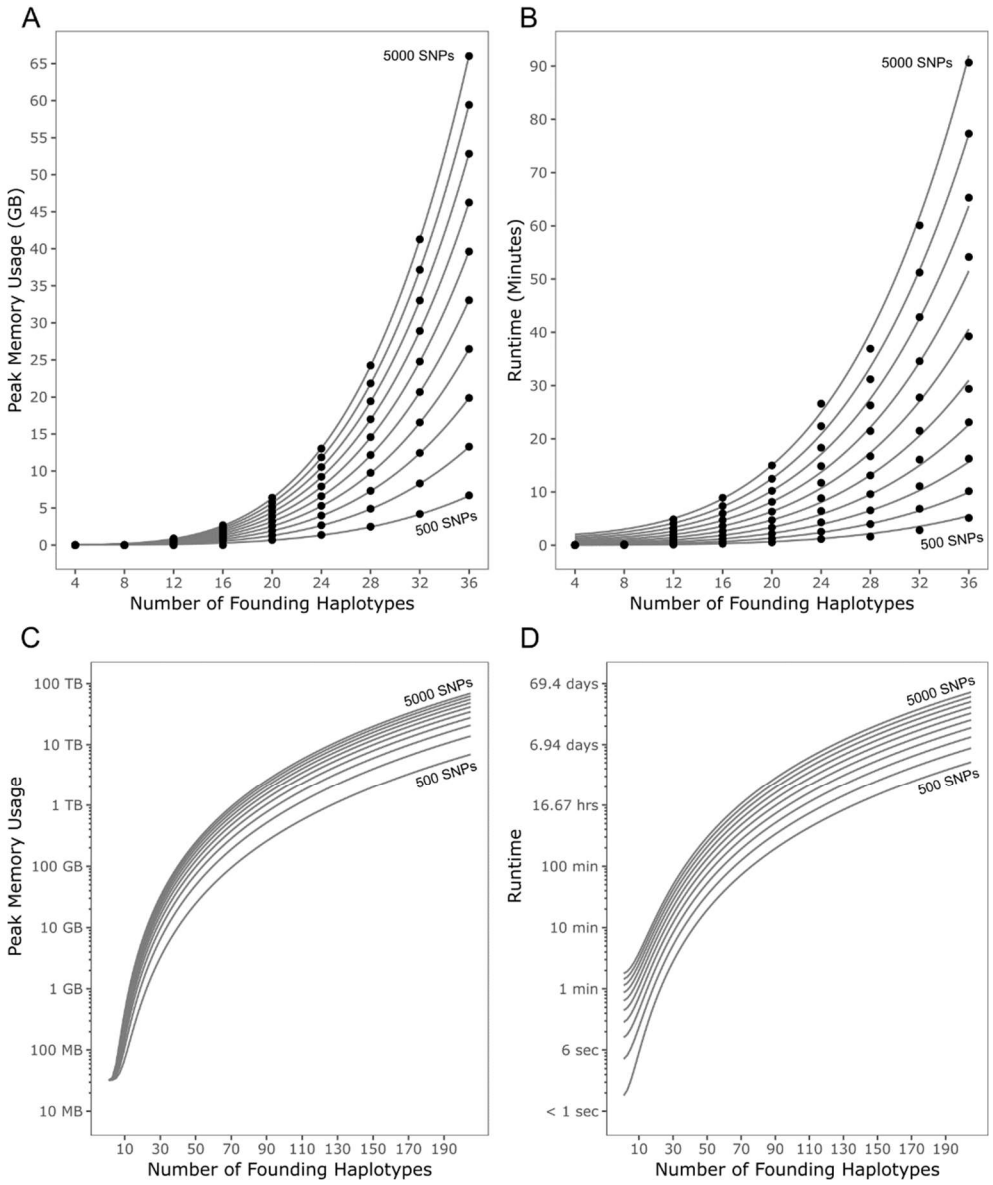


Figure 1. Resource usage of RABBIT during haplotype reconstruction. All reconstructions involve the same simulated 2L chromosome arm comprised of four haplotypes. Simulations included varied numbers of founding haplotypes (*N*) and a randomly selected set of markers (number of SNPs, *S*, incremented 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. Regression models for memory usage and runtime are extrapolated over a greater range for number of founding haplotypes in **C** and **D**, respectively.

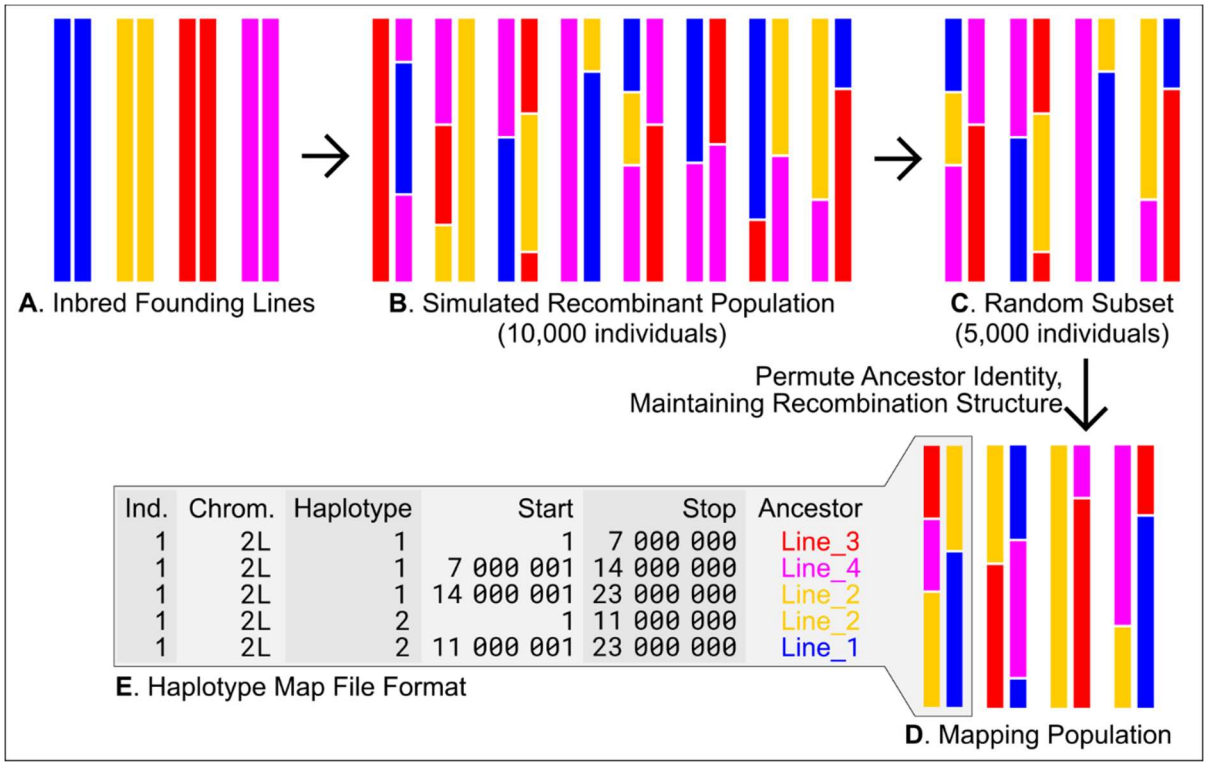


Figure 2. Basic structure of the forward simulator pipeline.
Inbred founding lines (**A**) are randomly intercrossed to produce a recombinant population (**B**). Rapid generation of independent mapping populations is achieved by random down-sampling (**C**) and permutation of ancestry (**D**). Population 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.

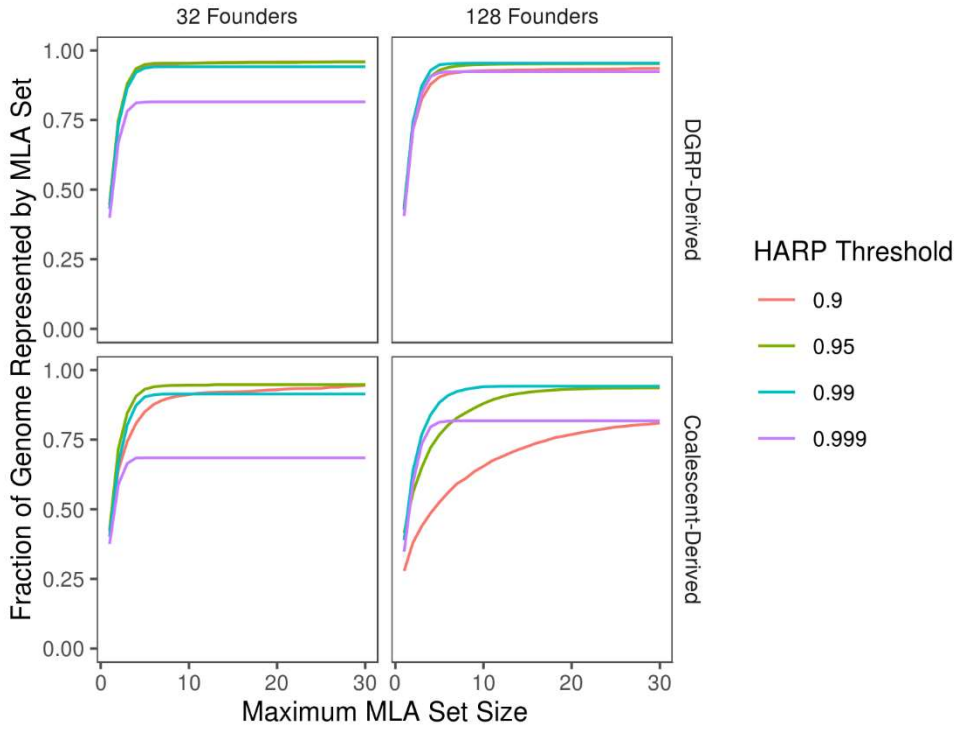


Figure 3. Optimization curves for Most-Likely-Ancestor (MLA) selection. To ensure reconstruction was both accurate and computationally feasible, we selected the smallest set of most-likely-ancestors that still represent the greatest fraction of to-be-reconstructed genomes. Increasing the upper limit for the number of MLAs chosen improves the fraction of genome represented with diminishing returns. The HARP threshold of 0.99 performed best for 128-founder populations, for both DGRP-derived and coalescent-derived simulations. Conversely, 32-founder populations performed best with a HARP threshold of 0.95. 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 $\mu = 5 \times 10^{-9}$.

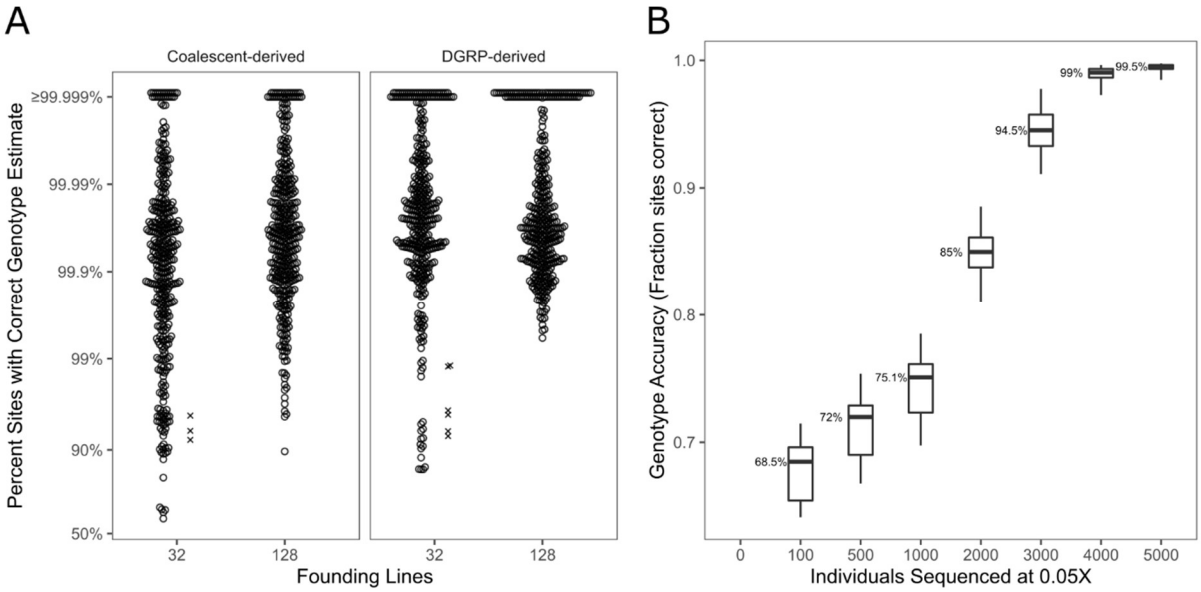


Figure 4. Genotype estimation accuracy for Hybrid Swarm populations. **A**: Reconstructions were performed for populations simulated as being founded by either 32 or 128 inbred lines at 0.05X sequencing coverage with up to 16 MLAs as determined with a HARP threshold of 0.99. Accuracy, calculated as the 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 (≥10 recombination events) displayed as an ×. 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 $\mu = 5 \times 10^{-9}$. **B**: Genotyping accuracy of STITCH for simulated 32-founder F_5 hybrid swarm populations sequenced at 0.05X coverage. Genotype accuracy improves with greater numbers of sequenced individuals, as STITCH infers missing genotypes from other haplotypes in the population. Boxes represent the median and interquartile range; whiskers extending to the lower and upper bounds of the 95% quantiles. See methods for parameters used.

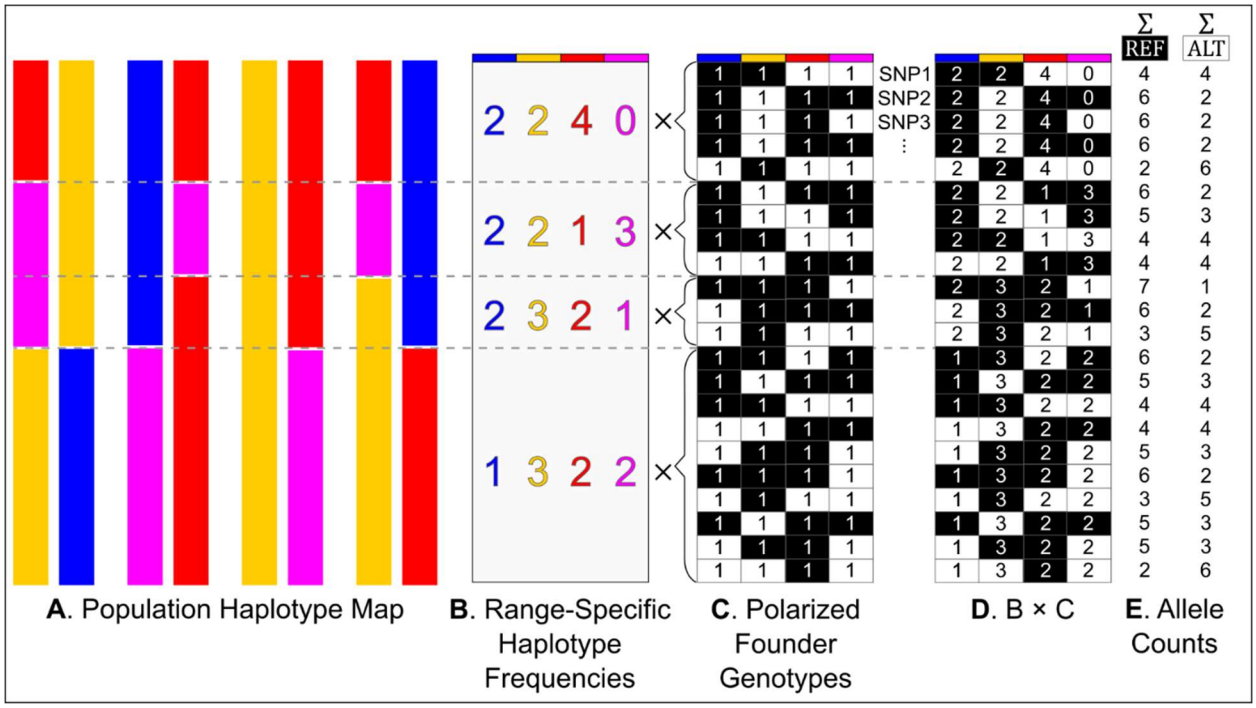


Figure 5. Schematic for rapid association testing with the haplotype block files. For a given population represented by a haplotype map file (**A**), all SNPs between sorted breakpoints (indicated by dashed lines) will share identical aggregated haplotype frequencies (**B**). Haplotype frequencies are 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.

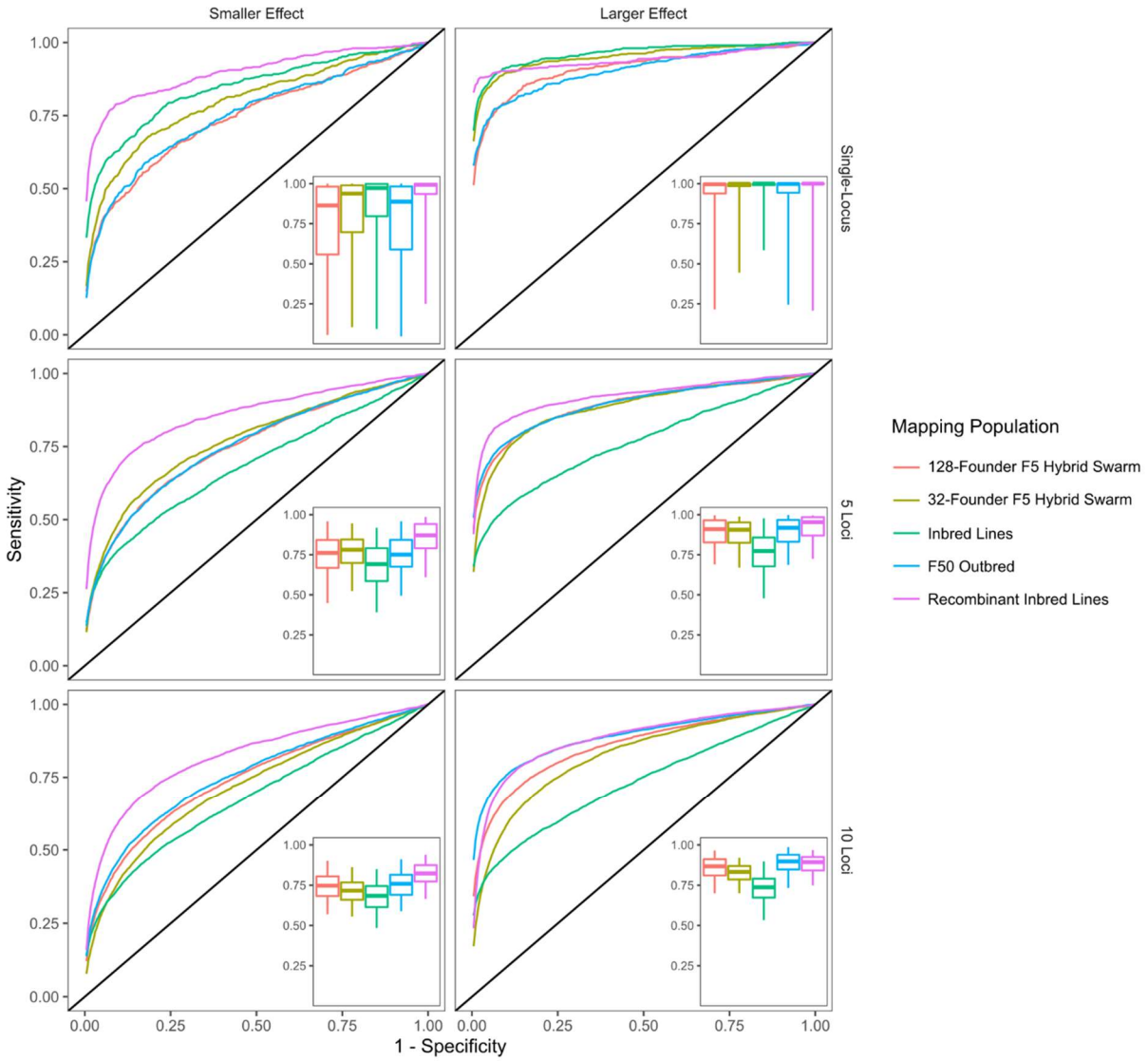


Figure 6. Receiver Operator Characteristic (ROC) curves for GWAS simulations. ROC curves represent the aggregate of 500 GWAS simulations per parameter combination. To generate single representative ROC curves across 500 GWAS simulations per population, we stepped through all specificity values and calculated the mean sensitivity. Inset boxplots display the distribution of AUCs, with whiskers spanning the middle 95% of data.

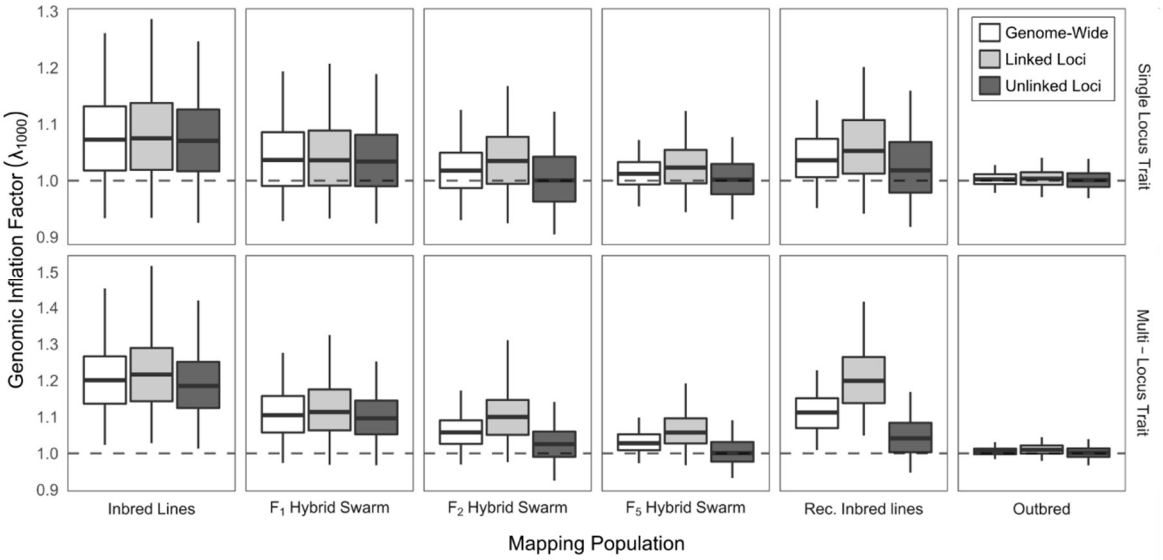


Figure 7. Genomic Inflation Factor (GIF, λ_{1000}) for simulated GWAS with monogenic trait of with a large effect locus, or multi-locus trait of smaller individual effects. GIF is calculated genome-wide (across all autosomes); on the autosome arm containing the causal allele (linked); and for sites on 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 sampled with replacement, phenotypically assigned in a case-control framework. Boxes represent the median and interquartile range; whiskers extending to the lower and upper bounds of the 95% quantiles. Hybrid Swarm populations here used 128 founders. **Outbred**: F_{50} population founded by 128 inbred lines.

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). The coalescent-derived populations described here were simulated with an effective population size of $N_e = 1 \times 10^6$ and mutation rate $\mu = 5 \times 10^{-9}$.

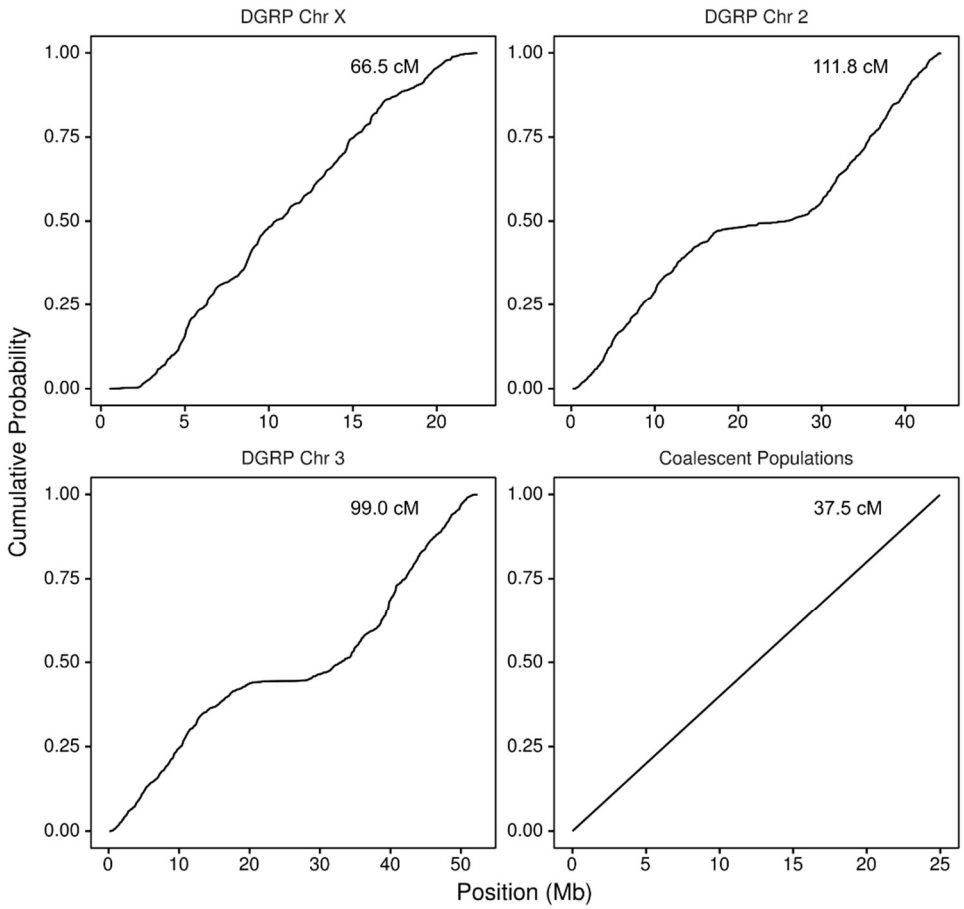


Figure S1. Recombination probability functions used for simulated individuals. Recombination is modeled as a Poisson process, with position sampled from linear interpolation of recombination rates measured in *Drosophila melanogaster* by Comeron et al. (2012). The frequency of recombination samples cumulative 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.

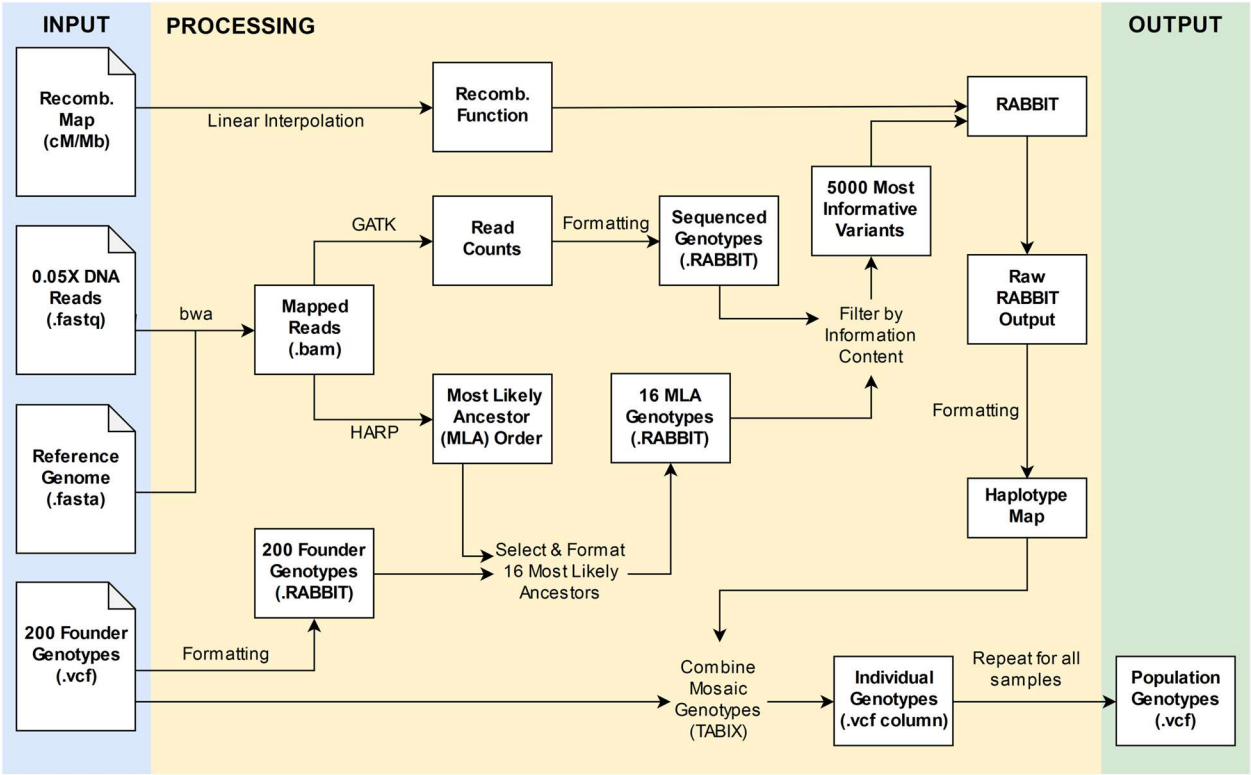
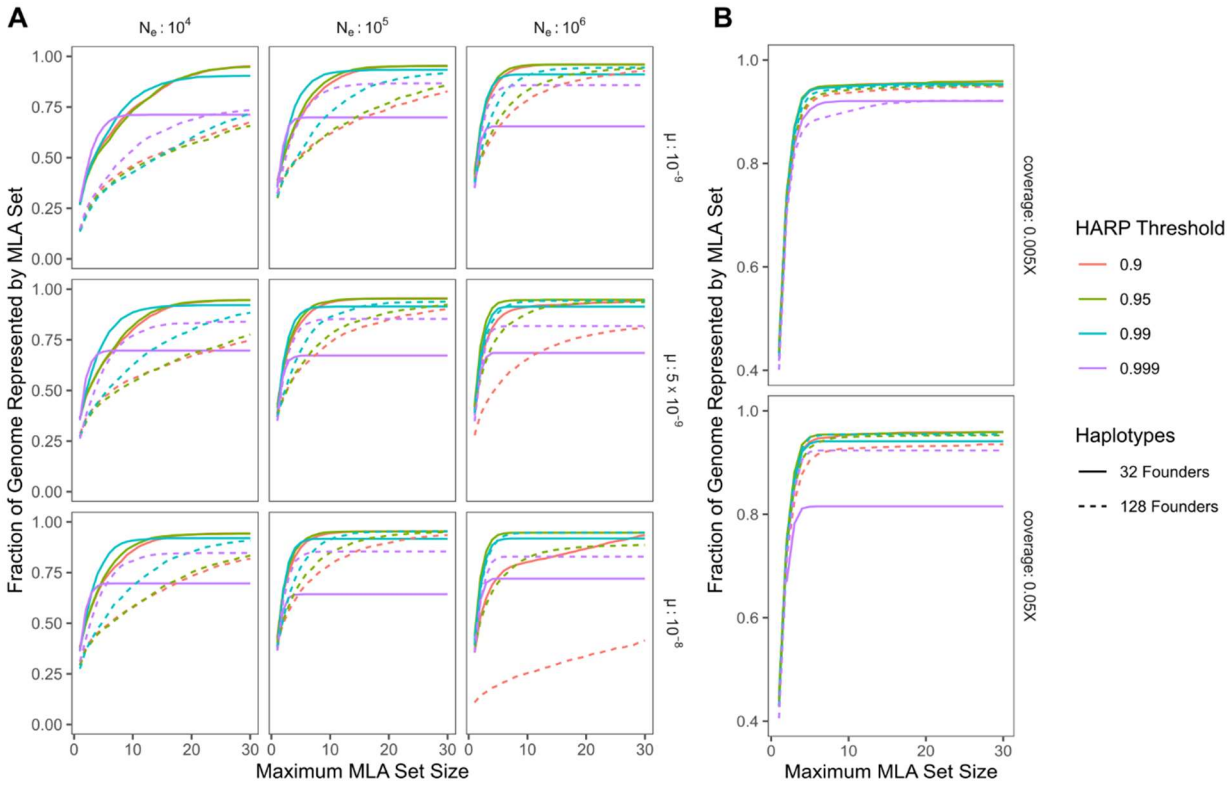


Figure S2. Workflow for reconstructing whole genomes from ultra-low coverage sequencing data.

Three required inputs are sequencing reads (.fastq), a reference genome (.fasta), and homozygous (or phased) founder haplotypes (.vcf file). A recombination map can optionally be provided for use with RABBIT to inform variation in recombination rates across chromosomes. Flowchart designed in draw.io.

1205



1206

1207 **Figure S3**. Optimization curves for Most-Likely-Ancestor inclusion for additional

1208 parameter combinations.

1209 A: SCRM (coalescent) derived populations across different parameters; B: accuracy is
1210 similar for both 0.05X and 0.005X sequencing coverage in populations simulated with
1211 DGRP genomes.

1212

1213

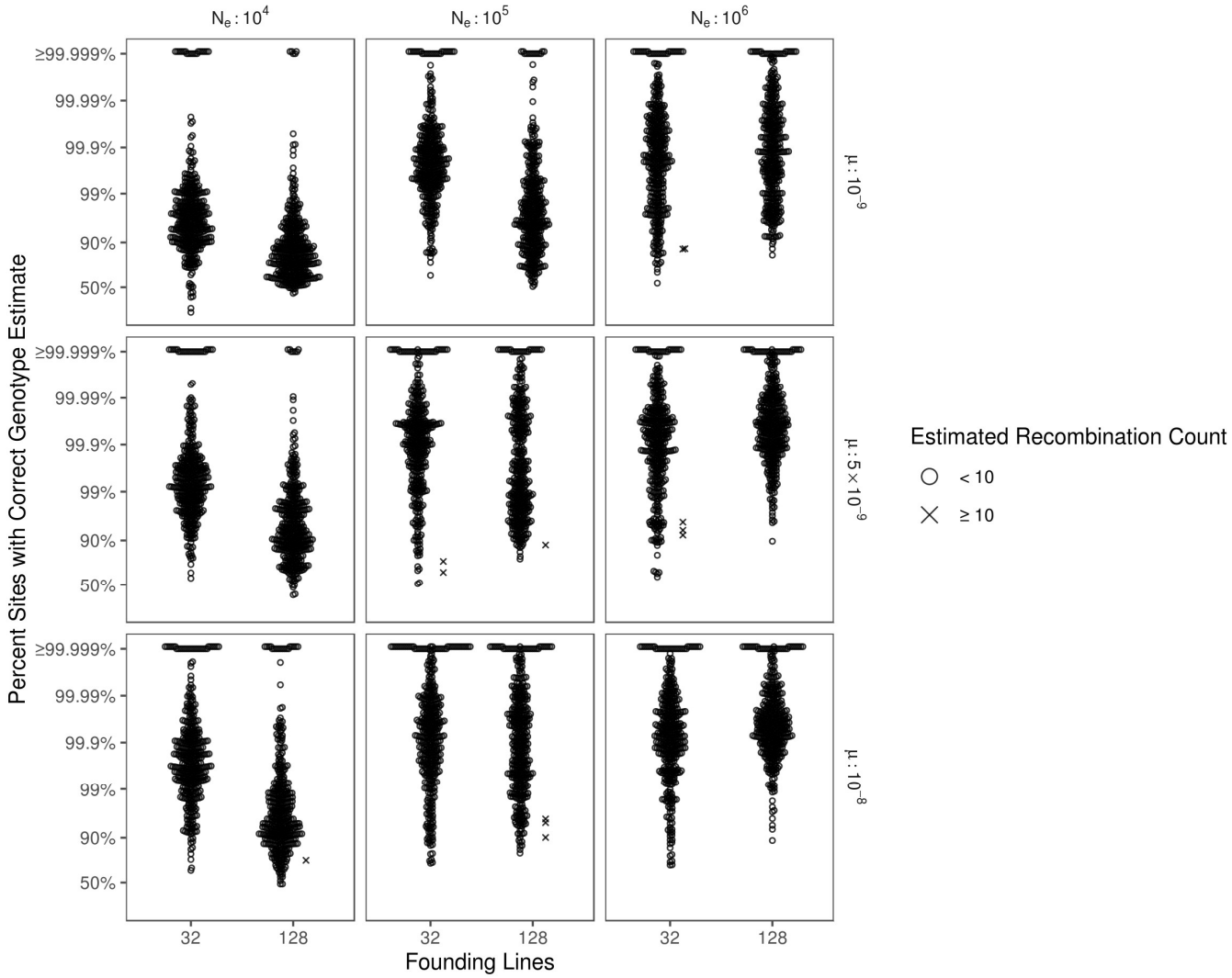
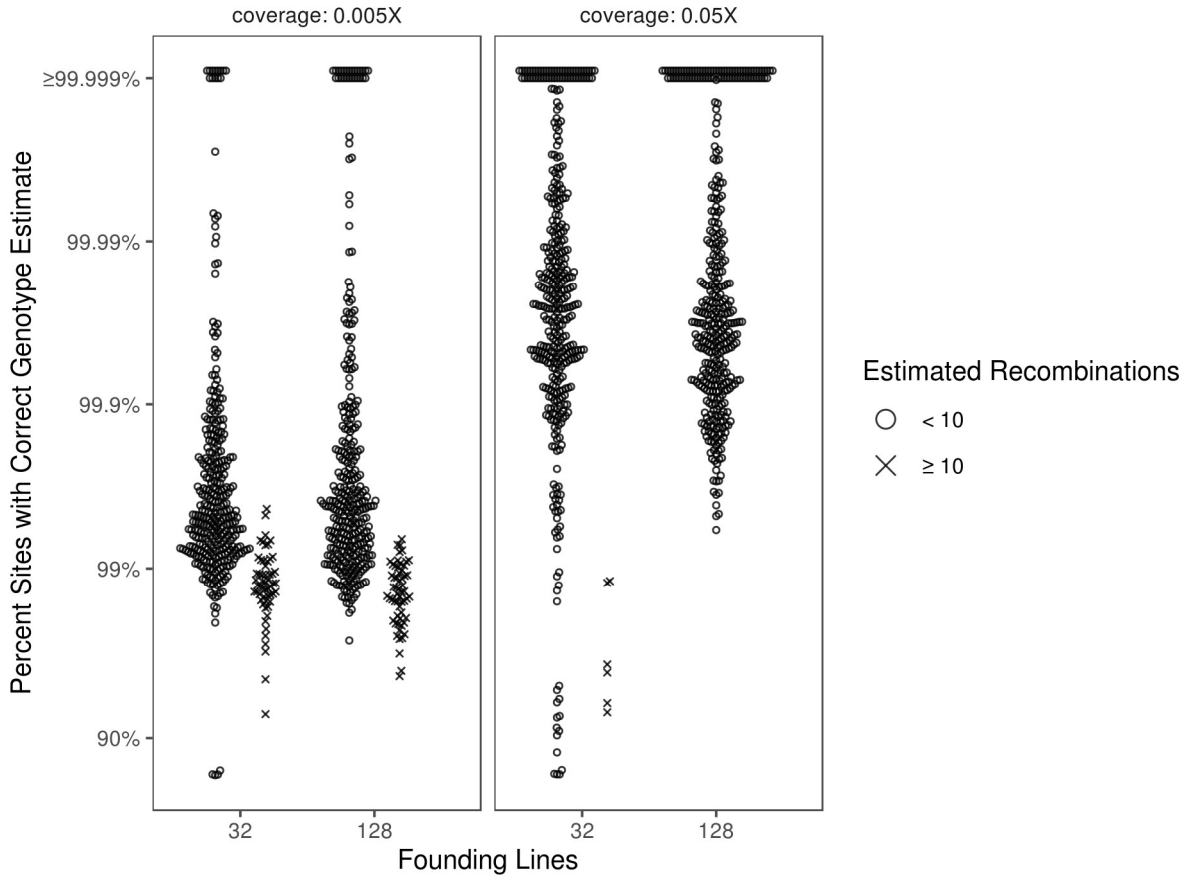


Figure S4. Accuracy of genome reconstruction for simulated, coalescent-derived F_5 hybrid swarm individuals.

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 estimated to have $\geq$10 recombination events are denoted by an ×, offset from the bulk of the distributions. Each parameter combination includes 400 reconstructed chromosomes (from 100 simulated individuals).

1224



1225

1226 **Figure S5.** Accuracy of genome reconstruction for simulated, DGRP-derived F_5 hybrid
1227 swarm individuals.
1228 Reconstructions were performed for populations simulated as being founded by either
1229 32 or 128 inbred lines for two levels of ultra-low sequencing coverage. Accuracy is
1230 represented on a logit scale, as most points occur above 90%. Reconstructed
1231 chromosomes estimated to have $\geq$10 recombination events are denoted by an ×, offset
1232 from the bulk of the distributions . Each parameter combination includes 400
1233 reconstructed chromosomes (from 100 simulated individuals).
1234

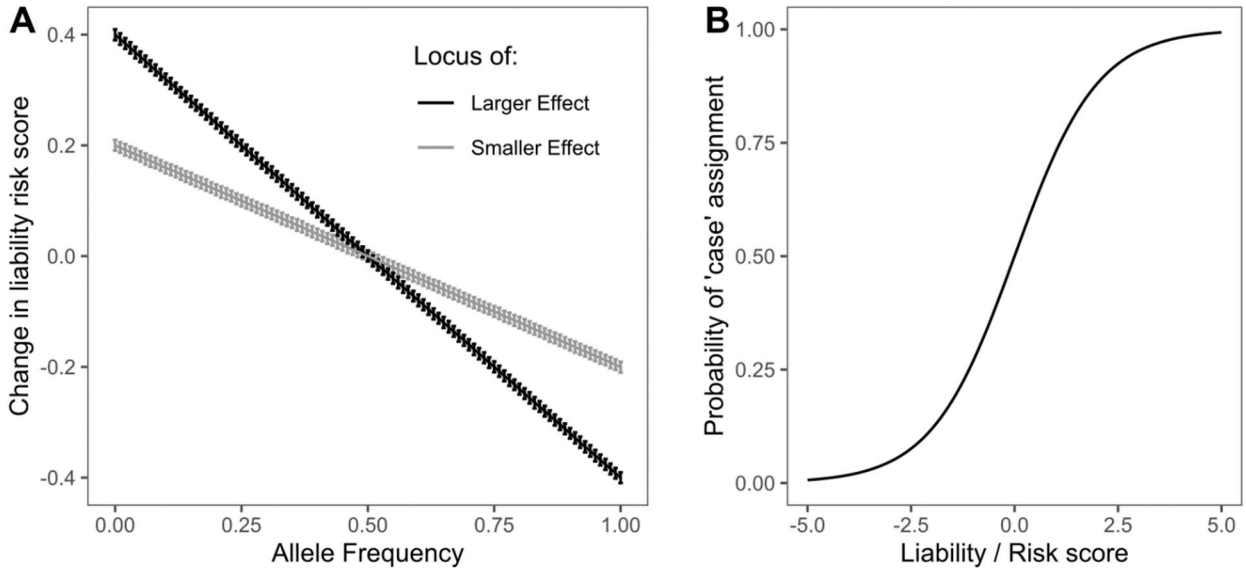


Figure S6. Liability phenotype model for GWAS simulations.

for transforming additive risk scores to probability of assignment to the case population, i.e. probability of developing a disease. Rare alleles increase risk score, which in turns increases probability of being assigned a case phenotype. A: The relationship between allele frequency and effect on liability risk score. Simulations were modeled as larger or smaller effect with a user-defined effect size coefficient. B: Logistic transformation of the form $y = (1 + e^{-x})^{-1}$ translates the liability score into probability of being assigned to case population. Most simulated individuals inhabit the linear portion of the curve, exemplifying additive genetics. Because phenotypic assignment is necessarily bounded by zero and one, simulated individuals that happen to inherit many loci of large effect will display antagonistic pleiotropy, where the combined effect of multiple loci is less than the sum of individual effects.

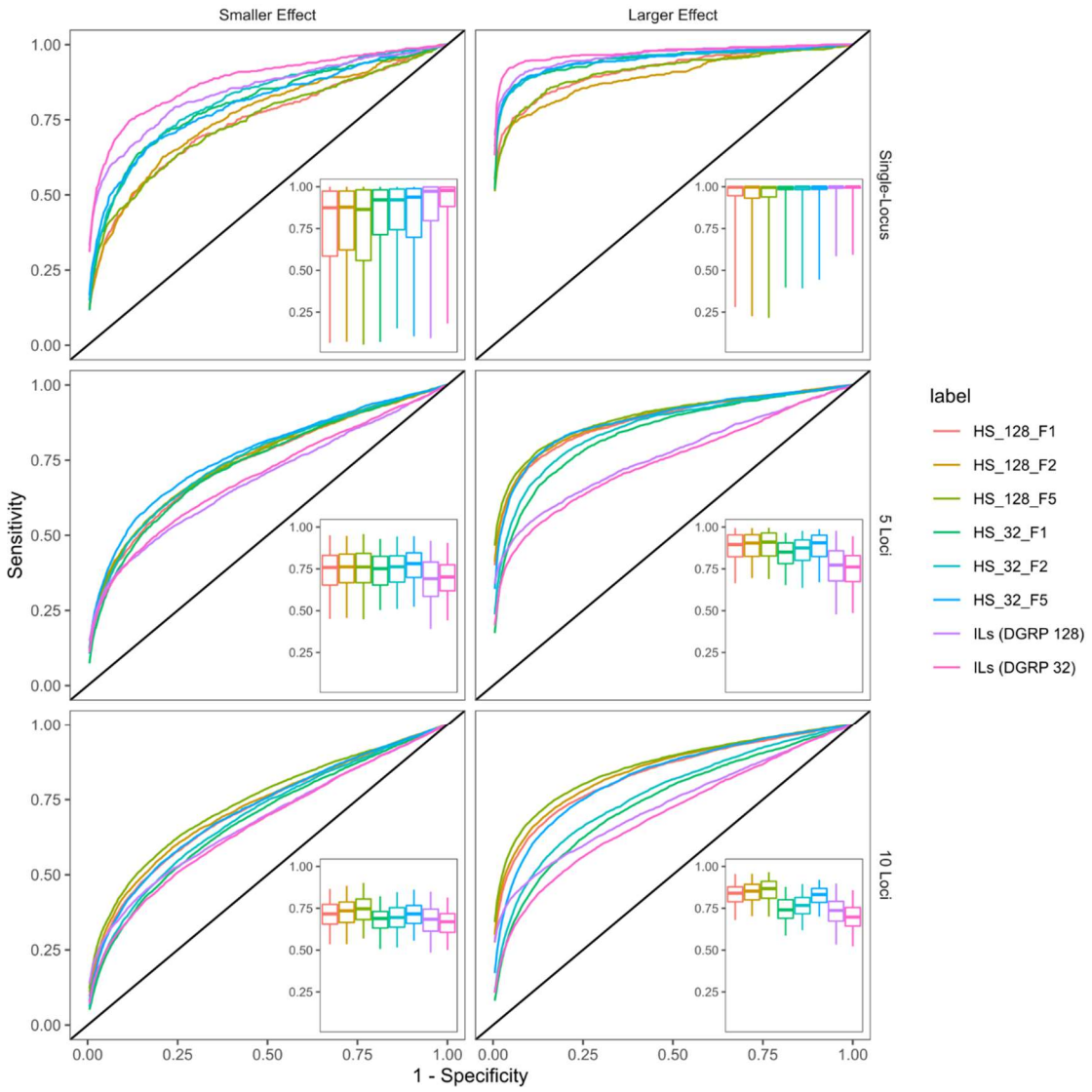


Figure S7. Receiver Operator Characteristic (ROC) curves for additional GWAS simulations.

To generate single representative ROC curves across 500 GWAS simulations per population, we stepped through all specificity values and calculated the mean sensitivity. Inset boxplots display the median and interquartile range of Area Under the Curve (AUC) distributions, with whiskers spanning the middle 95% of data. **ILs**: inbred lines. **HS**: Hybrid Swarm populations founded by 32 or 128 lines.

1259
1260

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

1261 **Table S1**. Accuracy of estimated number of recombination events following
1262 chromosome reconstruction.
1263 A high concordance correlation coefficient (Lin's ρ) indicates agreement between
1264 estimated and true recombination counts for 400 reconstructed chromosomes
1265 (coalescent-derived populations) or chromosome arms (DGRP-derived populations).
1266 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).