DOI: 10.1534/genetics.104.039313

Simulating the Collaborative Cross: Power of Quantitative Trait Loci Detection and Mapping Resolution in Large Sets of Recombinant Inbred Strains of Mice

William Valdar,[1] Jonathan Flint and Richard Mott

Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom

Manuscript received December 3, 2004
Accepted for publication December 7, 2005

ABSTRACT

It has been suggested that the collaborative cross, a large set of recombinant inbred strains derived from eight inbred mouse strains, would be a powerful resource for the dissection of complex phenotypes. Here we use simulation to investigate the power of the collaborative cross to detect and map small genetic effects. We show that for a fixed population of 1000 individuals, 500 RI lines bred using a modified version of the collaborative cross design are adequate to map a single additive locus that accounts for 5% of the phenotypic variation to within 0.96 cM. In the presence of strong epistasis more strains can improve detection, but 500 lines still provide sufficient resolution to meet most goals of the collaborative cross. However, even with a very large panel of RILs, mapping resolution may not be sufficient to identify single genes unambiguously. Our results are generally applicable to the design of RILs in other species.

THE Complex Trait Consortium (CTC) outlined a strategy, called the collaborative cross (CC), to construct a very large set of recombinant inbred (RI) strains from a genetically diverse set of inbred strains of mice (Churchill *et al.* 2004). It is suggested that the CC is an ideal resource for systems biology as it will provide a reproducible, highly varied yet controlled set of genetic backgrounds for functional genomic studies. One projected use of the CC is the detection and mapping of genetic effects underlying complex phenotypes. In principle, the CC could solve the obdurate problem of high-resolution quantitative trait locus (QTL) mapping: the CTC position paper suggested the creation of 1000 RIs to map QTL to a resolution where gene identification becomes possible, within a reasonable time and budget. An international collaboration to breed these lines has started and is described at http://www.complextrait.org.

However, the experimental design and statistical power of the CC has not yet been rigorously assessed. A number of factors need to be considered. First, the genetic architecture of complex traits, which remains largely unknown, will influence the CC's ability to detect and resolve genetic effects. Second, it is not clear how best to design the cross to maximize and maintain genetic diversity. If the CC is not seeded with sufficient diversity in its initial stages, or loses diversity during breeding, there may be insufficient detectable recombinants in the RI strains to deliver adequate mapping resolution. Third, assuming that the CC does have the potential to deliver high-resolution QTL mapping, how does it compare to other possibly more efficient strategies? For instance, detection and fine mapping can be achieved within a population of manageable size by making use of historical recombinants. Advanced intercross lines (AILs), populations made from F_2's by repeated intercrossing within a large population, have the simple genetic background necessary for unambiguous assignment of identity by descent along with the highly recombinant haplotype structure that allows high mapping resolution (Darvasi and Soller 1995; Wang *et al.* 2003). Heterogeneous stocks (HS) are a generalization of AILs, derived from a cross of eight inbred mouse lines (McClearn *et al.* 1970; Demarest *et al.* 2001). Although HS are more complex to analyze, they provide additional genetic diversity and may offer increased mapping resolution over AILs (Talbot *et al.* 2003). But AILs and HS have drawbacks, since they take years to breed and are expensive to genotype. The efficiency of AIL and HS designs, relative to the CC, has not been investigated.

In this article, we use simulation to compare QTL detection and accuracy of the collaborative cross design against several alternatives. We simulate the realistic situation of a quantitative trait that owes half its variance to genetic factors. Of that half, a small portion is due to a QTL on a genotyped chromosome. The remaining variance is attributable to multiple QTL that lie on ungenotyped, hence unobserved, chromosomes. We investigate the ability of the CC to detect and map a QTL when it acts independently and when it is under strong epistatic control. We investigate variations on the basic breeding design and compare it with traditional alternatives (F_2 intercross and the backcross) and newer,

[1]*Corresponding author:* Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Dr., Oxford OX3 7BN, United Kingdom. E-mail: valdar@well.ox.ac.uk

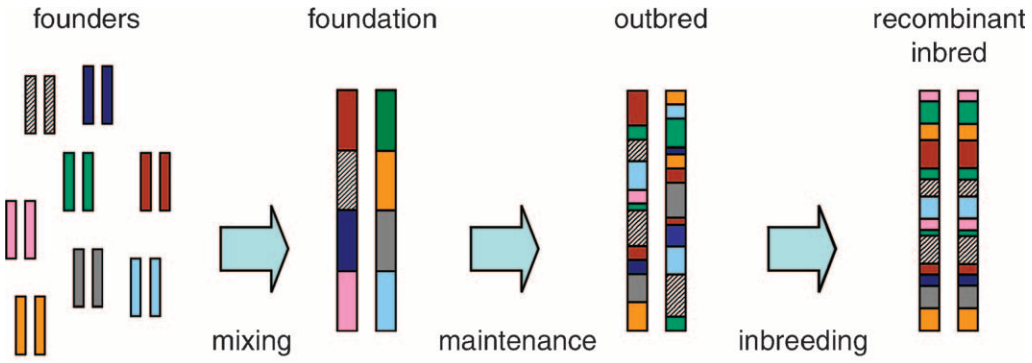

Figure 1.—Components of a generalized breeding program that leads to a mosiac cross (see methods). The first stage, mixing, combines different color-coded inbred strains to produce mice containing some genetic material from each founder. In the optional maintenance stage, mice are intercrossed to increase the number of recombination breakpoints. A subsequent inbreeding produces recombinant inbred lines.

related strategies [recombinant inbred heterogeneous stocks (RIHS) (Valdar *et al.* 2003) and RIs based on AILs]. Finally, we ask how many lines are necessary to fine map small-effect QTL both in simple additive and in strongly epistatic systems.

METHODS

Breeding strategies overview: We examined five breeding strategies: the CC (Complex Trait Consortium 2003; Churchill *et al.* 2004), RIHS (Valdar *et al.* 2003), recombinant inbred advanced intercross lines (RIAILs), the F_2 intercross (F_2), and the backcross (BC). RIHS and RIAILs are panels of recombinant inbred strains made from, respectively, HS and advanced intercross (AI) populations. The HS, RIHS, AI, RIAIL, and CC all share a common aim: to produce fine-grain haplotypic mosaics from a limited known genetic repertoire. For this reason, we term them "mosaic crosses." Moreover, with the exception of the backcross, the mosiac crosses and the F_2 are variants of a more general multistage breeding program. That program has three stages: a compulsory mixing stage, followed by two optional stages, maintenance and inbreeding (Figure 1). In the mixing stage, inbred progenitor strains are intercrossed to produce a foundation population of mice whose genomes each contain some genetic material from every founder strain. In the maintenance stage, foundation mice are intercrossed over a number of generations. The aim is to introduce recombinants and produce a population of mice whose genomes are fine-grained mosaics of the original founder haplotypes. Intercrossing in the maintenance stage is typically performed using a pseudo-random mating technique that avoids pairing of close relatives and thereby reduces genetic drift and the risk of fixation. In the inbreeding stage, randomly chosen pairs of mice are inbred by recurrent brother–sister mating over 20 generations, each pair producing a distinct recombinant inbred line. We describe the four main types of strategies, and their variants, below and in Table 1.

F_2 intercross and backcross: In an F_2 cross two inbred lines, A and B, are crossed in the mixing stage, to make F_1 foundation animals, of genetic composition AB. They enter a maintenance stage of one generation to produce F_2 animals. The inbreeding stage is omitted. For the backcross, two inbred lines, A and B, are combined in an F_1 cross. These foundation animals are then crossed with strain A to produce animals whose chromosome pair comprises one A haplotype and one recombinant AB haplotype.

RIAILs: Advanced intercross lines (Darvasi and Soller 1995) resemble an F_2 with an extended maintenance stage that lasts two or more generations. AIL genomes thus also have genetic composition AB, but contain more recombinants. To provide an informative comparison with the CC and RIHS crosses, which is the true focus of our work, we do not simulate AILs as such, but instead recombinant inbred strains derived from an AIL (RIAIL), made by adding on an inbreeding stage to the traditional AIL design. The genomes of RIAIL mice are therefore highly recombinant mosaics of strains A and B that theoretically are homozygous and identical within a particular line but different between lines. We use a notation in which RIAILs are suffixed with the number of filial generations intercrossed before inbreeding, *e.g.*, RIAIL25 means RIs produced from an F_{25} or, equivalently, an AIL25 population.

RIHS: Recombinant inbred heterogeneous stocks (Valdar *et al.* 2003) are like RIAILs but derived from eight founders rather than from two. They are made as follows. Eight inbred lines, A–H, are combined in a three-step mixing phase. In the first step, four F_1 populations are produced of compositions AB, CD, EF, and GH by mating A with B, C with D, and so on, *i.e.*, a "two-way" cross. The second step proceeds with two "four-way" crosses: AB and CD mice are mated to produce mice of composition ABCD, while EF and GH similarly give rise to EFGH. The last step is an "eight-way" cross, in which ABCD mice are crossed with EFGH mice to produce mice of genetic composition ABCDEFGH. We call these animals foundation heterogeneous stocks, or HS0. After maintenance for n generations to build up recombinants they become HSn. Subsequent inbreeding produces RIHSn; *e.g.*, RIHS30 are RIs produced by inbreeding pairs of HS30 mice.

TABLE 1

Strategies simulated

Type of strategy	Strategy name	Mixing	No. of maintenance generations	Inbreeding for 20 generations	Total no. of generations
CC	cc	Disjoint diallel	0	yes	3 + 0 + 20 = 23
CC	cctop10	Disjoint diallel[a]	0	yes	3 + 0 + 20 = 23
CC	cc04	Disjoint diallel	4	yes	3 + 4 + 20 = 27
CC	cc08	Disjoint diallel	8	yes	3 + 8 + 20 = 31
RIHS	rihs05	Simple	5	yes	3 + 5 + 20 = 28
RIHS	rihs10	Simple	10	yes	3 + 10 + 20 = 33
RIHS	rihs30	Simple	30	yes	3 + 30 + 20 = 53
RIAIL	riail08	Simple	7	yes	1 + 7 + 20 = 28
RIAIL	riail13	Simple	10	yes	1 + 10 + 20 = 33
BC	bc	Simple	1[b]	no	1 + 1 + 0 = 2
F_2	f2	Simple	1	no	1 + 1 + 0 = 2

CC, collaborative cross; RIHS, recombinant inbred heterogeneous stocks; RIAIL, recombinant inbred advanced intercross lines; F_2, F_2 intercross. See METHODS for descriptions of "disjoint diallel" and "simple" mixing.

[a] Strategy includes selection: only the most highly recombinant 10% of animals from the four-way (F_2) cross participate in the eight-way (F_3) cross.

[b] Maintenance involves mating with one of the parental lines in the case of BC.

CC: The collaborative cross as initially proposed (COMPLEX TRAIT CONSORTIUM 2003; CHURCHILL *et al.* 2004) is much like RIHS but with a more elaborate mixing stage and no maintenance breeding. In its most extreme form, mixing is ambitiously complex. Simpler mixings have since been suggested (CHURCHILL *et al.* 2004), whose complexity lies between that of the original and that of simple RIHS mixing. Here we simulate the most extreme form, noting that this may overestimate the efficacy of an intermediate. As in RIHS, the mixing stage of the collaborative cross has three steps. The first step produces F_1's from every distinct pair of founders, generating mice of composition AB, AC, . . . , FH, GH. This "diallel" design comprises 28 two-way crosses or 56 if reciprocal crosses between males and females are performed (which they are not here as we restrict our attention to autosomes). In step two, every possible four-way cross is performed between F_1's, excluding instances of shared descent. For example, AB × CD is allowed but AB × BC is not. We call this a "disjoint diallel." In the last step, eight-way crosses are performed in a further disjoint diallel (where, for example, ABCD × EFGH is allowed but ABCD × ABDE is not), resulting in foundation mice of composition ABCDEFGH. Its repeated disjoint diallel structure means the mixing stage incurs a combinatorial explosion of crosses. In the unrealistically conservative situation of every distinct cross being performed with no more than one mating pair, step two would see 210 matings and step three, 310. Only a subset of those eight-way crosses would be feasible in practice. We simulate them all.

In the standard CC proposal, pairs of foundation mice then are inbred to produce RIs. We also investigated variant CC designs with more recombination, to increase mapping accuracy. In "cctop10," we select only the most highly recombinant 10% of mice from the four-way cross (which could be determined by genotyping) to proceed to the eight-way cross. In "cc04" and "cc08," we insert four and eight maintenance generations (∼1 and 2 years of extra breeding), respectively, between the mixing and inbreeding.

Maintenance stage: There are several established ways to maintain a laboratory population over many generations with a view to preserving heterozygosity (CABALLERO and TORO 2000). Circular mating (also known as rotational breeding) defines a structured mixing of the animals, repeated at each generation that implicitly restricts the consanguinity of each mating and so minimizes the extent of inbreeding, drift, and fixation (KIMURA and CROW 1963; BOUCHER and COTTERMAN 1990). We perform maintenance in all strategies using circular mating. We simulate it as follows. Given a population of $2m$ individuals, split equally between the sexes, put couples into cages labeled 1–m. Prepare empty cages labeled $1'$–m' to hold the next generation. Mate the female from cage 1 with the male from cage 2 and send their offspring to cage $1'$. In similar fashion, mate the female of cage k with the male from cage $k + 1$ (or the male from cage 1 if $k = m$) and send their offspring to cage k', doing this for all remaining k. Repeat this procedure, with suitable relabeling, until the required number of maintenance generations is reached.

The genetic diversity of the final population depends critically on m, the number of mating pairs. When m is small, the maintenance stage acts as a population bottleneck that encourages extremes of drift and fixation.

Conversely, a larger m delays these effects. Consistent with existing HS mice (Northport HS) (Demarest *et al.* 2001), we set m at 50 when generating RIHS lines unless otherwise specified and do this for RIAILs also. However, for our CC variants, a project of more ambitious scope, we set m equal to the number of RIs required, which can be as high as 1000.

Founder strains and marker sets: Northport HS stocks derive from a cross of the mouse strains A/J, AKR/J, BALB/cByJ, C3H/HeJ, C57BL/6J, CBA/J, DBA/2J, and LP/J. R. Hitzemann (personal communication) chose those strains with dual purpose: to emulate the original Boulder HS, whose founders were genetically diverse by the standards of that day, and to provide a contrast of response classes to haloperidol-induced catalepsy.

The Complex Trait Consortium (2003) urges that the founder set chosen for the CC balances genetic diversity, phenotypic diversity, breeding performance, and the availability of complementary resources. The strains finally selected for breeding by the CTC were A/J, C57BL/6J, 129Sv/ImJ, NOD/LtJ, NZO/H1J, CAST/Ei, PWK/Ph, and WSB/Ei, which include a number of wild-derived strains. However, at the time of simulation this set was not yet finalized, and only incomplete genotype data were available for them. For these reasons, and because we wished to compare alternative eight-founder designs on an equal footing, all simulated populations were founded with the Northport set listed above or, in the case of the two-strain strategies, with pairs of strains drawn from that list. The genomes of wild-derived strains such as CAST/Ei are more distantly related than are standard strains such as C57BL/6J. Therefore we would expect more QTL to segregate in the actual CC but the results of our simulations should not change overmuch since they are conditioned on the presence of segregating QTL.

We use two marker sets. The first is an artificial fine-grain set of 1000 fully informative markers (*i.e.*, markers with eight alleles) spaced at regular intervals of 0.1 cM. We use this set only to establish unrecombined block length for different designs (see later) and not for mapping. The second, which we use to map QTL, is a real data set typed for chromosome 1, provided by the Genomics Institute of the Novartis Research Foundation (GNF) (Pletcher *et al.* 2004). We selected markers that segregate among the eight Northport HS progenitors to make our simulations as realistic as possible. The GNF set numbers 512 diallelic SNPs with average spacing 0.19 cM (standard deviation 0.16 cM) over 100 cM. These markers segregate among the eight founders with varying minor allele frequencies. The strain distribution pattern of the SNPs is such that of the 512 minor alleles, 68 are present in only one strain, 136 are present in two strains, 209 in three strains, and 99 in four strains.

Representing the mouse genome: We represent a mouse by up to seven pairs of homologous autosomes, depending on the complexity of the genetic model being simulated. Chromosome 1 is 100 cM long and contains the GNF marker set or in the case of establishing block lengths the fine-grain artificial marker set. The QTL to be mapped always lies on chromosome 1. The remaining six chromosomes act as an unseen genetic background that affects the phenotype; these chromosomes are ignored during mapping. For our simulations of epistasis, chromosome 2 contains a single diallelic marker that acts as a masking QTL in our simulation of epistasis but is ignored otherwise. Chromosomes 3–7 each contain 10 diallelic markers spaced at 10-cM intervals. These 50 markers are used as background QTL (see below). Each of the 51 markers on chromosomes 2–7 has its high allele on four founders independently and randomly chosen from the eight.

Simulating breeding in the mouse population: We simulate mouse populations stochastically, using software written in house (available from http://www.well.ox.ac.uk/~valdar/software/). Gametogenesis is simulated by randomly drawing one of each parental autosome and recombining it with its homolog, choosing breakpoints according to the Haldane model with no interference, in which meioses occur as events in a one-dimensional Poisson process (Lynch and Walsh 1998).

Phenotypic variance components: We simulate a quantitative phenotype that is 50% heritable in the founder population. This means that genetic factors explain half the phenotypic variance and environmental noise accounts for the remainder. The heritable portion of the variance is split between foreground and background QTL. Foreground QTL are defined as those that lie on chromosome 1. These QTL are visible in our mapping. Background QTL lie on the unseen chromosomes (2–7) and, although they affect the phenotype, are unmappable. The phenotypic variance partitions thus,

$$\begin{aligned}\sigma^2_{\mathrm{P}} &= \sigma^2_{\mathrm{G}} + \sigma^2_{\mathrm{E}} \\ &= \left(\sigma^2_Q + \sigma^2_B\right) + \sigma^2_{\mathrm{E}},\end{aligned}$$

where subscripts identify phenotypic (P), environmental (E), genetic (G), foreground (Q), and background (B) variance components, and where, for a phenotype that is 50% heritable, $\sigma^2_{\mathrm{G}} = \frac{1}{2}\sigma^2_{\mathrm{P}}$.

QTL effect sizes and allelic values: The effect size of a QTL is defined here as the proportion of phenotypic variance it explains in the founder population, *e.g.*, $\theta_Q = \sigma^2_{\mathrm{G}}/\sigma^2_{\mathrm{P}}$. We explicitly vary the effect size of a single foreground QTL while implicitly adjusting the effect sizes of the background QTL to make up the remaining heritability. These 50 background QTL act independently to provide only additive genetic variation. Thus, ignoring covariance between QTL (which is negligible in our design), $0.5 = \theta_Q + \sum_{i=1}^{50} \theta_{B_i}$, where the effect size of the ith background QTL is $\theta_{B_i} = (0.5 - \theta_Q)/50$, such that in any one simulation all background QTL are equally potent in the founder population. The case of epistasis is described below.

The allelic value is defined here as the raw number that a QTL's high allele contributes to the phenotype value. We adjust each QTL's allelic value at the start of each simulation to yield the required effect size in the founder population. That allelic value then remains constant throughout breeding and phenotyping. This means, owing to inevitable drifts in allele frequency, that a QTL's effect size in the mapping population may differ from its starting value in the founder population, but it models reality better than fixing the effect size in the mapping population by changing its allelic value.

Simulating QTL: We simulate two main foreground QTL systems: a single additive QTL and a single QTL in strong epistasis with an unseen QTL on chromosome 2. In both cases, background QTL are present but act independently of QTL in the foreground.

We simulate a single foreground QTL as follows. Before any breeding takes place, a marker is drawn at random from chromosome 1 and designated the QTL. Then the strain distribution pattern (SDP) of its alleles is shuffled. We do this for the following reason. In the GNF marker set, 90% of the SNPs segregate between C57BL/6 and A/J or DBA/2, due to the method of SNP ascertainment. As a result, the strain distribution pattern of the QTL matches that of nearby SNPs more frequently than would be likely in reality, since there is no reason to suppose a QTL favors any particular SDP. To correct for this bias, we shuffle the SDP of the QTL but keep the SDP of the remaining SNPs unchanged. For example, a QTL with its minor allele in strains A, B, and C may, after shuffling, find that allele moved to strains B, F, and H. Were the strain distribution pattern of the QTL not shuffled it would be easy to detect through single-marker association because its allelic state would be closely matched by those of its flanking markers. We demonstrate this point by performing an additional set of unshuffled simulations for one experimental condition, contrasting its results with those of the shuffled case (see RESULTS). Once phenotyping is complete, the foreground QTL marker is concealed, and we then attempt to recover that marker's location.

The high allele of each QTL is assigned an allelic value on the basis of its frequency and required effect size in the founder population. The phenotype of each animal in the mapping population is calculated as $y = \alpha_Q x_Q + \sum_{B=1}^{50} \alpha_{B_i} x_{B_i} + \varepsilon$, where, with subscripts B identifying background loci, α_B denotes the allelic value, x_B the number of high alleles, and ε a normally distributed error term with variance accounting for half the total variance in the founders. Table 2 illustrates the relation between the genotype (x) of an additive locus and its genetic effect (αx). For inbred mice x is usually 0 or 2 (although it can be 1 if inbreeding has not fixed that QTL); for F_2 mice x is 0, 1, or 2; and for BC mice x is 0 or 1.

For simulations of epistasis, the foreground and background QTL are generated as before except that the genetic effect of the foreground QTL on chromosome 1 now depends on the masking QTL, *M*, on chromosome 2. The equation for a mouse phenotype is now $y = \alpha_{Q\times M} x_M x_Q + \sum_{i=1}^{50} \alpha_{B_i} x_{B_i} + \varepsilon$, such that, for example, in a population of inbred lines the genetic effect from the foreground QTL ($\alpha_{Q\times M} x_M x_Q$) is nonzero only when both foreground and masking QTL are in the high state. Note that the "effect size" of the foreground QTL is now defined as the proportion of variance explained by the QTL pair. The goal of subsequent detection and fine mapping is to find the first QTL despite the other's confounding effect.

The foreground and masking QTL always segregate in each simulation. Thus for F_2, RIAIL, and BC, founded on just two strains, the pair of strains chosen is random, subject to the constraint that those QTL segregate between them.

Detection and fine mapping: We detect and fine map simulated QTL using two methods, single-marker association (SMA) and HAPPY, a multipoint method (MOTT *et al.* 2000). In single-marker association, we regress the phenotype on the number of minor alleles present at each marker in turn. If the marker showing the strongest association is statistically significant (see later), we classify the QTL as detected and predicted to be located in the most significant marker interval. Otherwise, we deem the QTL undetected.

Mapping using HAPPY is also performed in a regression framework, but here the phenotype is regressed on the inferred probability of descent at each marker interval. Briefly, HAPPY models genotypes as symbols

TABLE 2

Mean effects for each multilocus genotype in simulated additive and epistatic QTL systems

QTL system	Genotype	Genetic effect (arbitrary units)
Single	aa	0
	aA	5
	AA	10
Epistatic	aabb	0
	aAbb	0
	AAbb	0
	aaBb	0
	aABb	5
	AABb	10
	aaBB	0
	aABB	10
	AABB	20

In nonepistatic simulations, the genetic effects of background and foreground QTL follow the additive scheme. In epistatic simulations, background QTL act additively but the foreground QTL and the unlinked masking QTL act according to the epistatic scheme. Note that, although listed, heterozygous genotypes are extremely rare in inbred RIAIL, RIHS, and CC populations. For this reason, we excluded dominance effects from either model.

emitted from hidden progenitor states in a discrete Markov process where transitions between states correspond to recombination events. This hidden Markov model (HMM) is fitted to the data using a dynamic programming algorithm described in Mott *et al.* (2000). The HMM transition matrix is parameterized by the expected number of recombinants between markers, which, given the centimorgan distance between markers, can be summarized by the effective number of generations g_e. Inbreeding, among other things, causes g_e to deviate from the actual number of generations and so g_e varies with breeding strategy. For each strategy we estimate g_e by calibrating the observed level of informative recombination against that seen in a simulated outbred population.

By combining all the genotype data along a diploid chromosome, HAPPY computes for each individual and at each locus the probability of descent from each pair of the founder strains. The progenitor probabilities are regressed on the phenotype to identify loci where the pattern of strain inheritance correlates with the phenotype. As with SMA, the marker interval that associates most strongly with the phenotype is tested for statistical significance and, if it passes, declared the putative QTL.

In each simulation, 1000 animals are generated. As a consequence, in some experiments multiple animals are phenotyped from each RI strain. In that case we regress the mean phenotype of those replicates (*i.e.*, the "strain mean") on genetic information from a randomly chosen member. Residual heterozygosity within an RI means that "replicates" are rarely identical at all loci, so we implicitly model the drawbacks of minimal genotyping.

Assessing statistical significance using generalized extreme value distributions: We measure the association between locus and phenotype in both single-marker association and HAPPY by an *F*-test of their respective linear model fits. The overall score for a simulation over m loci is defined as follows. Let $S_i = -\log_{10} P_i$, where P_i is the *P*-value from the *F*-test at locus i. Define $S_{\max} = \max\{S_1, \ldots, S_m\}$, that is, the maximum score over all loci. The reported *P*-value for the highest-scoring locus is grossly anticonservative. We therefore assess statistical significance by comparing the obtained *P*-value with genomewide thresholds that have been corrected for multiple testing. The Bonferroni correction is too conservative as it ignores linkage disequilibrium (LD) between markers, and a permutation test is computationally unfeasible for so many simulations. Instead, we take LD into account by modeling the distribution of $S_{\max}$ under the null hypothesis of no foreground QTL and hence calculate suitable threshold quantiles.

The LD structure varies with breeding strategy, number of RI lines, and extent of background genetic variation so we estimate the null distribution of $S_{\max}$ separately for each experimental condition. Provided there is negligible long-range dependence between distant S_i, with increasing m, $S_{\max}$ asymptotically follows a generalized extreme value distribution (Coles 2001) with distribution function $G(S_{\max}) = \exp\{-[1 + \xi(S_{\max} - \mu)/\sigma]_+^{-1/\xi}\}$, where μ, σ, and ξ are the location, scale, and shape parameters, respectively, and $h_+ = \max(h, 0)$. A closely related use of generalized extreme values (GEVs) is described in Dudbridge and Koeleman (2004).

For each experimental condition we perform 1000 null simulations with background QTL only. We then fit a GEV by maximum likelihood to the observed values of $S_{\max}$, using the R package "evd" (Stephenson 2003). An example of fit is shown in Figure 2a, which plots

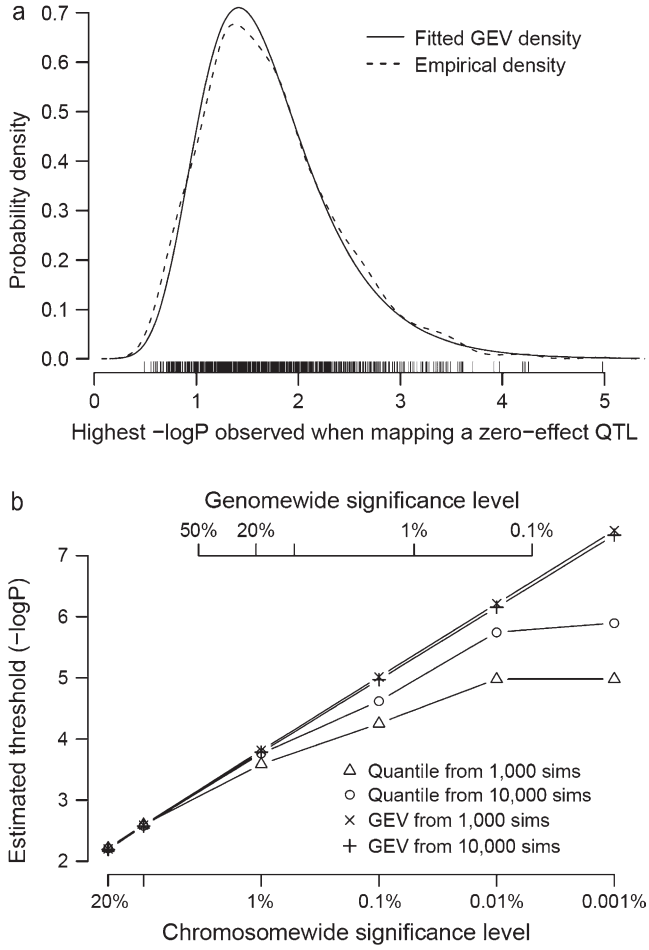

Figure 2.—Fitting a generalized extreme value (GEV) distribution to the null distribution of maximum $-\log P$ for the cc strategy with 500 RILs. (a) The empirical and fitted densities for $-\log P$ of the best-fitting locus in each of 1000 simulations of a zero-effect QTL. Underneath the curves, a rug plot shows individual data points. (b) Chromosomewide and genomewide thresholds estimated by two methods: "Quantile" takes a suitable quantile of the empirical distribution (*e.g.*, the 5% chromosomewide threshold would equal the 95% quantile of the distribution of observed scores), whereas "GEV" estimates thresholds analytically via a fitted GEV. Estimates are plotted for 1000 and 10,000 null simulations, with 10,000 expected to be more accurate. The plot shows that whereas quantile estimates vary substantially with the number of simulations and especially for small significance levels, GEV estimates are relatively consistent.

TABLE 3

Significance thresholds for HAPPY scores (*i.e.*, $-\log_{10}$ *F*-test *P*-values) calculated for different combinations of breeding strategy and number of RILs

	Estimated 5% genomewide significance threshold (observed quantile in 1000 null simulations)		
Strategy	0 RILs	500 RILs	1000 RILs
cc[a]	—	4.51 (4.69)	4.49 (4.19)
cc04	—	4.75 (5.22)	4.59 (4.41)
cc08	—	4.88 (4.36)	4.95 (4.58)
cctop10	—	4.65 (4.81)	4.4 (5.05)
riail08	—	6.2 (6.46)	6.96 (7.02)
riail13	—	7.39 (8.55)	8.83 (9.87)
rihs05	—	9.15 (9.52)	13.93 (14.15)
rihs10	—	14.33 (12.61)	22.75 (19.48)
rihs30	—	28.22 (19.95)	41.49 (29.77)
BC	5.03 (4.57)	—	—
F_2	5.2 (5.08)	—	—

Each threshold is calculated from a generalized extreme value (GEV) distribution that has been fit to the maximum scores obtained from 1000 simulations of mapping a zero-effect QTL (*i.e.*, 1000 null simulations). Null simulations listed here had the level of background genetic variation at 45% and so apply only to mapping of a 5% foreground QTL. In parentheses are thresholds calculated directly from the empirical cumulative distribution of null maxima (shown for comparison). Note that correction for anticipated multiple chromosomes means the 5% genomewide level corresponds to a chromosomewide significance level of ~0.26%.

[a] Additional simulations were performed for cc, with results [written as number of RILs = power (location error)]: 100 RILs = 4.62 (4.49), 200 RILs = 4.62 (5.25), 250 RILs = 4.49 (4.35), 333 RILs = 4.43 (4.55).

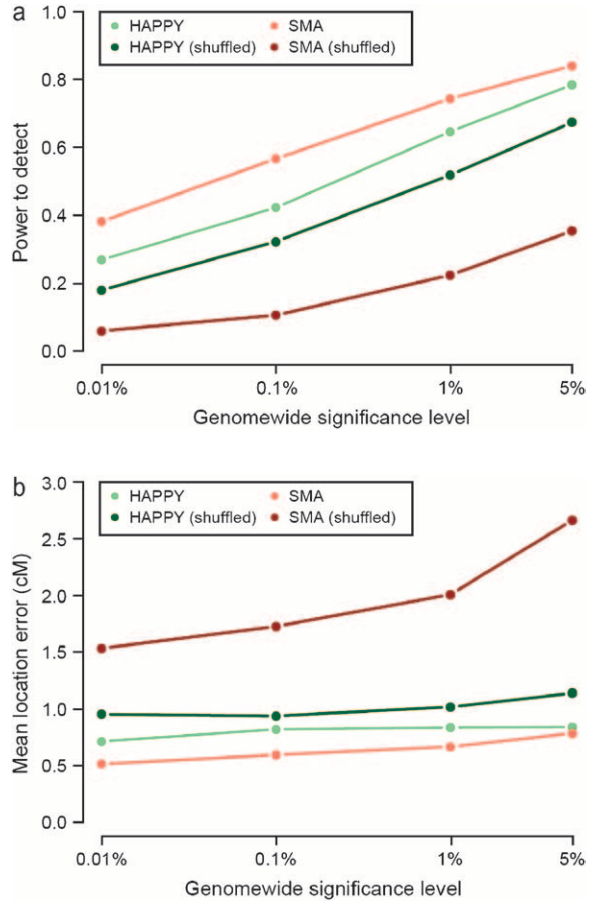

FIGURE 3.—The effects of marker ascertainment bias on mapping power and resolution. A diallelic 5% QTL was mapped in a population of 1000 mice derived from 500 cc RIs, using either the original strain distribution pattern (green) or a shuffled SDP (red) for two mapping algorithms SMA and HAPPY. Shuffling removes any marker ascertainment bias. (a) Power to detect a significant QTL. (b) The location error for a predicted QTL.

empirical and fitted distributions of S_{max} for the cc strategy with 500 recombinant inbred lines (RILs). To determine genomewide significance thresholds we used the upper tail quantiles corresponding to *P*-values of $g(5\%)$, $g(1\%)$, $g(0.1\%)$, and $g(0.01\%)$, where $g(x) = 1 - (1 - x)^{1/20}$ Bonferroni corrects for the fact that we simulate only one chromosome per animal but set thresholds as if there were 20 [*i.e.*, $g(5\%) \approx 0.26\%$].

Table 3 lists these significance thresholds calculated for all combinations of strategies and numbers of RILs for the case of a total background effect of $\theta_B = 45\%$.

We found that, for a given number of null simulations, estimating thresholds via a GEV model is more efficient than the traditional approach of taking empirical quantiles from the observed cumulative distribution of S_{max}. To illustrate this point, Figure 2b shows how threshold estimates from both methods vary with the number of null simulations available.

Assessing the efficacy of detection and fine mapping: For a given breeding protocol and QTL system, we calculate two main quantities from 1000 independent simulations: power to detect and location error. "Power to detect" is the probability that a QTL is detected at a given genomewide significance level α, defined as the proportion of trials in which the highest-scoring locus exceeds the α-threshold. "Mean location error" is the mean absolute distance of the predicted QTL from the actual QTL in trials where the predicted QTL reaches α-level significance. We investigated measuring accuracy using root mean squared error but found that its sensitivity to outliers (*i.e.*, false positive QTL detection events) produced misleading results.

RESULTS

Single-marker association compared to HAPPY for detecting QTL: We first investigated the effects of algorithm choice (SMA *vs.* HAPPY) and SNP ascertainment. Figure 3, a and b, plots the performance of HAPPY and single-marker association in fine mapping a 5% QTL

segregating in a population of 1000 mice derived from 500 collaborative cross RIs (*i.e.*, with 2 replicate mice per RI strain). Note that here we define a 5% QTL as one that accounts for 5% of the phenotypic variance among the eight founder strains and that we include additional unobserved QTL to make up the additive genetic variance to 50% (see METHODS).

Figure 3a shows the power to detect a significant QTL. The orange and light green lines show the performance of SMA and HAPPY in the case where the QTL is an unaltered SNP drawn from the original GNF marker set. The dark red and dark green lines show what happens when the SNP chosen to be the QTL has had its SDP shuffled with respect to the founders in advance of the simulation and represent a more likely scenario (see METHODS for details). In the first case, SMA does better than HAPPY, detecting the QTL with higher power at all thresholds. But if the SDP of the QTL is shuffled, then single-marker association fails whereas HAPPY remains powerful. Similarly, Figure 3b, which plots the location error against the significance threshold, tells a similar story: when the QTL is an unaltered SNP, both HAPPY and single-marker association predict its location with subcentimorgan accuracy. But when the alleles of the QTL have been shuffled, the location error of SMA predictions rises dramatically whereas those of HAPPY are almost unchanged. Hereafter, we consider only shuffled QTL.

Figure 4 plots the location error of HAPPY predictions against those of SMA in 1000 simulations. For F_2 populations (Figure 4a) the two methods are similar, with poor localization (where the peak is >15 cM away from the true location) slightly more in HAPPY than in SMA (3.1% *vs.* 1.2% of cases). However, for CC populations (Figure 4b) HAPPY localizes the QTL as accurately as before but SMA misses the QTL by >15 cM far more often (2.8% *vs.* 10.1% of cases). Hereafter, we consider only mapping predictions made by HAPPY for all strategies except BC and F_2, for which we consider predictions made by both HAPPY and SMA.

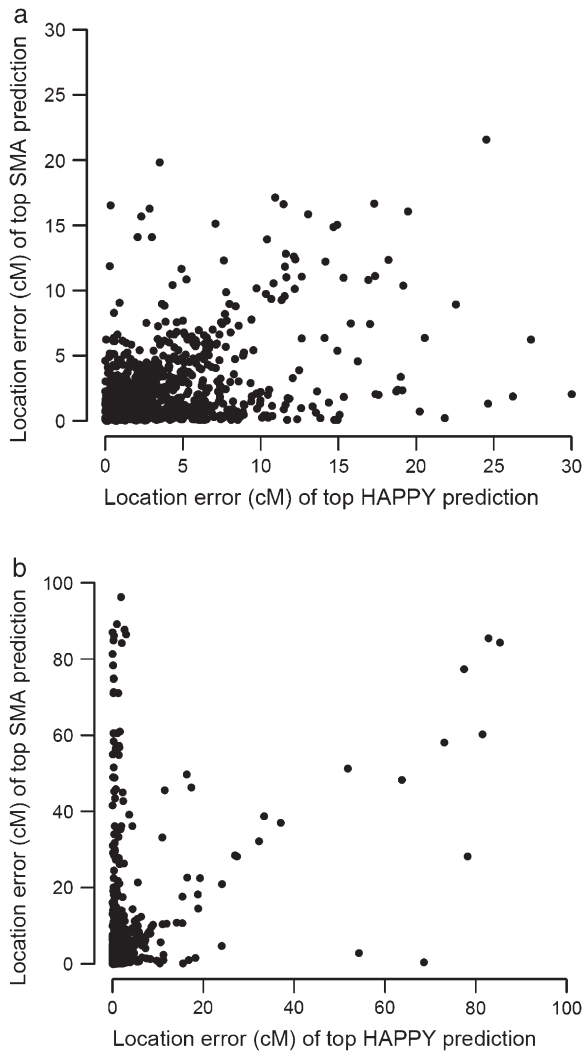

FIGURE 4.—Location error of predicted QTL using HAPPY *vs.* single-marker association (SMA) in 1000 simulated mapping populations. (a) One thousand F_2. (b) One thousand individuals from 500 collaborative cross recombinant inbred lines.

Variation in haplotype block length: Table 4 gives summary statistics of distributions of the lengths of unrecombined haplotype blocks produced by the breeding strategies. Statistics were calculated from 1000 simulations of 1000 animals, using the artificial fully informative marker set described in METHODS. The estimates have a slight upward bias because double recombinants returning to the same founder that occur between markers evade detection. Despite this, the observed distributions were close to the exponential distribution predicted by theory. As expected, the block length shortens with more generations. For instance, within the group of CC strategies, cc, which has no selection or maintenance, has the longest blocks whereas cc08, which has 2 years of maintenance, has the shortest. Strategies rihs05 and rihs10 have block lengths comparable to those of cc04 and cc08, and this owes much to their similar maintenance periods. Although the maintenance periods of cc04 and cc08 are also close to those of riail08 and riail13, these RIAIL strategies have longer block lengths because, with two founder haplotypes rather than eight, double recombinants are always invisible and many recombinations are uninformative. The shortest mean block length is seen in rihs30 at 4 cM. However, the block lengths of cc04 and rihs10 are not much longer, indicating that maintenance beyond a couple of years is subject to diminishing returns.

Single QTL—detection and resolution: We simulated a single additive QTL in RI, backcross, and F_2 intercross populations using a fixed population size of 1000 animals. The allelic value of the QTL was constant across designs and accounted for 5% of the phenotypic variance among the eight founders. For the RI designs, we varied the number of recombinant inbred lines from which the 1000 animals were drawn. For example, 100

TABLE 4

Unrecombined block length

Strategy name	Apparent unrecombined block length (cM)	
	Mean	SD
cc	11.51	12.18
cctop10	10.09	10.46
cc04	8.56	8.89
cc08	6.91	7.12
rihs05	8.66	9.05
rihs10	6.85	7.11
rihs30	4.00	4.15
riail08	12.90	13.66
riail13	10.47	11.14
BC	68.41	36.88
F_2	36.84	26.79

Mean and standard deviation (SD) for each strategy are estimated from 1000 simulated populations of 1000 individuals. The standard errors for the block length means are of the order 0.01.

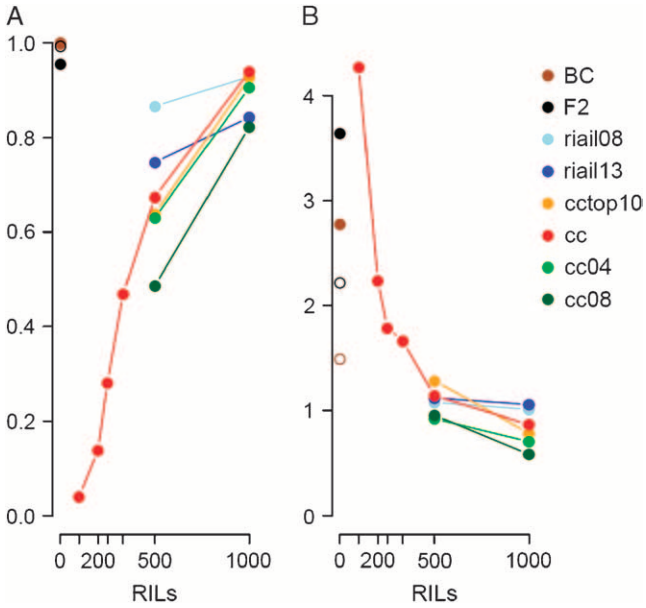

FIGURE 5.—Detection and mapping of a single additive 5% QTL using HAPPY (solid circles) and SMA (open circles, F_2 and BC only) for different breeding strategies. Each data point is based on 1000 simulations of mapping in 1000 individuals. The horizontal axes indicate the number of RI lines in the cross. Zero represents backcross (BC) and F_2 strategies; 100 (middle tick) refers to RI strategies with 10 replicates for each of 100 RI lines; 1000 means one mouse for each of 1000 RI lines. Note that a 5% QTL in an RI translates to a 2.56% QTL in an F_2 and a 1.3% QTL in a BC (see text). (A) Power to detect, *i.e.*, the probability that the highest-scoring putative QTL achieves 5% genomewide significance. (B) The mean location error (in centimorgans), *i.e.*, the mean absolute discrepancy between the actual and predicted location of the QTL, for mapping experiments that achieved 5% significance or better.

RILs correspond to 10 replicates from each of 100 lines. The results are plotted in Figure 5 and tabulated in Table 5.

Figure 5A shows the power to detect an effect at the 5% genomewide significance level plotted against the number of RILs (with non-RI strategies at 0 RILs). Power is defined here as the proportion of 1000 trials in which the highest-scoring locus exceeds a given significance threshold. The cc strategy (red line) illustrates the trend among RI strategies: the greater the number of lines is, the more power to detect. There is consistent variation between RI strategies too: at 500 RILs, two-way strategies (*i.e.*, those involving only two progenitor strains; blue) are more powerful than eight-way ones, and strategies with more recombination (*e.g.*, cc08 or riail13) do worse than their less recombinant counterparts. At 1000 RILs, the eight-way strategies catch up. Overall, BC and F_2, two-strain strategies with minimal recombination, have the greatest power to detect QTL.

Figure 5B shows the mean location error for detected QTL. Whereas each point in Figure 5A was based on 1000 simulations, in Figure 5B each point is based on only those simulations that were significant. For instance, the estimate for cc at 100 RILs (leftmost red point) is an average over only 40 simulations and so will have greater variance than the other plotted points. Nonetheless, the cc strategy (red) clearly shows how accuracy improves with increasing numbers of RILs. In particular, its elbow bend at 500 lines (1.1 cM) characterizes a process of diminishing returns where, past that point, modest gains hardly justify additional RILs. The remaining trends are the reverse of Figure 5A: more strains contributing to the cross and greater recombination give rise to sharper localization of the QTL, and this is most salient at 1000 RILs, where the best performer, cc08, reaches 0.58 cM. F_2 and BC deliver more accurate localization with SMA (open circles) than with HAPPY (solid circles) but nonetheless lag behind the ≥500 RIL strategies.

Figure 5 does not show results from the RIHS strategies because their power was too low to allow meaningful estimates of location error. Echoing the trend above, power and degree of recombination were inversely related with 1000 RILs of rihs05 detecting 14.1% of QTL, rihs10 detecting 2.2%, and rihs30 detecting 0.5% (see Table 5). The explanation appears to be genetic drift.

Genetic drift of the foreground QTL: By definition the QTL segregates in the progenitor strains and has an allele frequency of 0.125, 0.25, 0.375, or 0.5, depending on how many founders carry its high allele. During breeding, the frequency drifts unpredictably, affecting the QTL effect size and detectability in the final mapping population. Table 6 describes the genetic drift of the QTL under detection for the different breeding strategies and for different starting allele frequencies.

TABLE 5

Detection and mapping of a single 5% foreground QTL in 1000 animals

	Power to detect QTL			Mean location error (cM)		
Strategy	0 RILs	500 RILs	1000 RILs	0 RILs	500 RILs	1000 RILs
cc[a]	—	0.67	0.94	—	1.14	0.87
cc04	—	0.63	0.91	—	0.92	0.71
cc08	—	0.49	0.82	—	0.96	0.58
cctop10	—	0.64	0.93	—	1.28	0.79
riail08	—	0.87	0.93	—	1.08	1.01
riail13	—	0.75	0.84	—	1.12	1.06
rihs05	—	0.19	0.14	—	2.20	0.89
rihs10	—	0.04	0.02	—	3.39	8.50
rihs30	—	0.00	0.01	—	—	1.18
BC	1.00	—	—	2.78	—	—
BC (SMA)	1.00	—	—	1.49	—	—
F_2	0.96	—	—	3.64	—	—
F_2 (SMA)	0.99	—	—	2.22	—	—

See Figure 5 legend and RESULTS for details.

[a] Additional simulations were performed for strategy cc with results [written as number of RILs = power (location error)]: 100 RILs = 0.04 (4.27), 200 RILs = 0.14 (2.24), 250 RILs = 0.28 (1.79), 333 RILs = 0.47 (1.66).

We do not list the mean of the QTL allele frequency in the simulated mapping populations, since an average hides the underlying fluctuation. Rather, we give the standard deviation of allele frequencies, listing strategies prone to wide fluctuations first. The order of strategies closely matches the reversed order of power to detect a 5% QTL. Notably, rihs30 is the only breeding strategy where the allele frequency is so unstable as to drift to fixation occasionally.

QTL effect size: We investigated the ability of the cc strategy with 500 RILs to map QTL of greater effect. Figure 6 shows detection and mapping of QTL with effect sizes 5, 10, 20, 30, and 40% using both HAPPY and SMA. In all cases, heritability was set at 50% with background QTL making up the remaining genetic variance. Figure 6a shows that effect sizes of ≥10% are uniformly detected with HAPPY (solid circles). With SMA (open circles), guaranteed detection occurs for effect sizes of ≥30%. The HAPPY line in Figure 6b is asymptotic, with the mean localization error at 0.24 cM for a 40% QTL (note: the average spacing between markers in the marker set is 0.19 cM). The localization by SMA shows small but consistent improvement with QTL of effect size ≥10%, yet never falls below the 2-cM level.

TABLE 6

Genetic drift of the foreground QTL in different 500-RIL breeding strategies (with F_2 and BC included for comparison)

	Standard deviation of QTL minor allele frequency for starting frequency q				
Strategy	$q = 12.5\%$	$q = 25\%$	$q = 37.5\%$	$q = 50\%$	% fixed
rihs30	9.5	11.0	10.2	8.7	0.9
rihs10	6.9	8.8	8.1	6.7	0
rihs05	6.2	7.4	7.7	5.5	0
riail13	—	—	—	5.3	0
riail08	—	—	—	4.5	0
ccall2yr	1.5	2.1	2.3	1.4	0
ccall1yr	1.4	1.9	2.1	1.3	0
cctop10	—	—	—	1.2	0
ccall	—	—	—	1.2	0
BC	—	—	—	0.9	0
F_2	—	—	—	0.7	0

Before breeding, the QTL has a starting minor allele frequency of q in the founders. During breeding that frequency fluctuates and in the mapping population it may be elevated or depressed. The extent of drift from founder to mapped individual is shown, expressed as the standard deviation of allele frequency in the mapping population. The more drift there is, the greater the tendency for alleles to approach fixation.

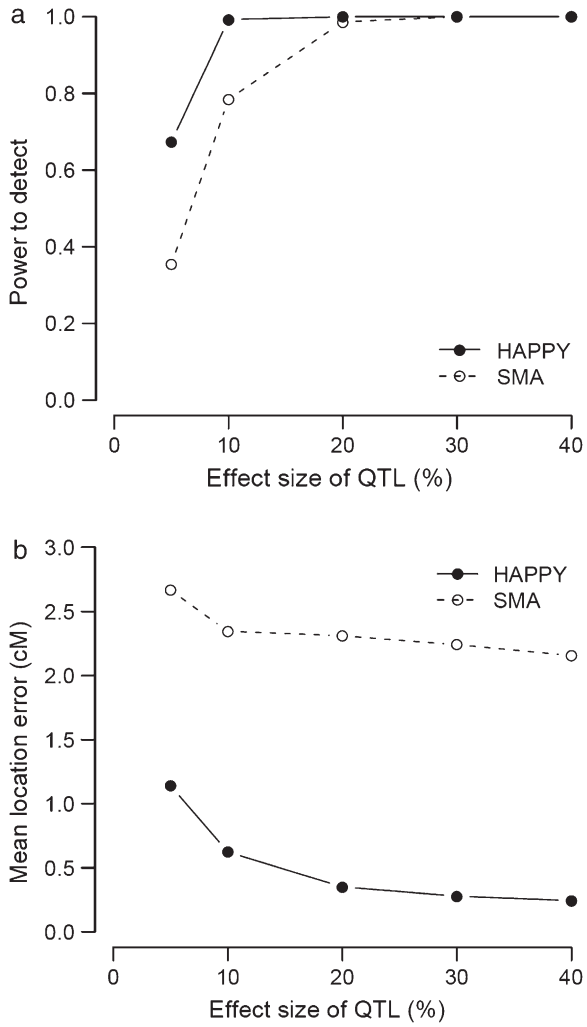

FIGURE 6.—Effect of varying QTL effect size on mapping strategy cc with 500 RILs. Results are shown for detection (a) and mapping (b) of a single additive QTL using HAPPY in a population of 1000 mice derived from 500 collaborative cross (strategy cc) RI lines. Values are plotted for effect sizes 5, 10, 20, . . . , 90%.

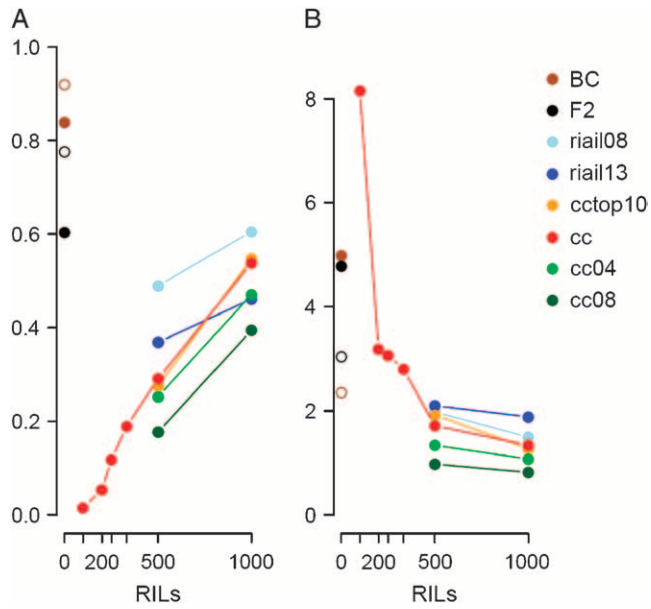

FIGURE 7.—Detection and mapping of an epistatic QTL. (A) Power to detect. (B) Mean location error (in centimorgans). Graphs show results of finding a single QTL embedded within an epistatic system in which the second QTL is invisible, unlinked, and has a masking effect and where the broad sense heritability of the two QTL is 10% (adjusted to 5.3% in F_2 and 2.8% in BC to maintain constant allelic value). Results are shown for HAPPY (solid circles) and SMA (open circles, F_2 and BC only) for different breeding strategies. Keeping the population size at 1000 but increasing the number of RI lines leads to increased power and accuracy. We illustrate this more fully for the cc strategy, reporting results for 100, 200, 250, 333, 500, and 1000 RI lines.

Epistasis: We next simulated a foreground QTL under strong epistatic control from an unlinked masking QTL lying on an ungenotyped chromosome (see METHODS) and tried to detect and map that foreground QTL in a population of 1000 animals. The combined effect size of the foreground and masking QTL was 10% in the founder population, with background QTL making up the genetic variance up to 50%. The results are plotted in Figure 7 and tabulated in Table 7.

Figure 7A shows the power to detect for the different breeding strategies. It exhibits the same trends as the detection graph in Figure 5 but with all points depressed by ~40%. The top performer is BC using SMA, which achieves >90% power. The RI strategies lag far behind, with the best two-way RI (riail08) reaching just above 60% and the best eight-way strategies (hardly distinguishable cc and cctop10) coming close to 55% at 1000 RILs. As before, the power of the RIHS strategies was too low to merit their inclusion in Figure 7, with 1000 RILs of rihs05, rihs10, and rihs30 achieving 9, 1.9, and 0.5%, respectively. Figure 7B plots location error and tells a similar story to that of Figure 5B, but with all strategies locating the QTL with roughly twice the error. RI strategies with ≥500 RILs outperform F_2 and BC strategies, although in the case of BC mapped with SMA this is by a small margin. The elbow bend of the cc line (500 RILs, 1.7 cM) shows, as before, that increasing the number of lines beyond 500 delivers only modest gains in resolution. The most accurate strategy and the only one that reaches below the 1-cM mark is cc08 (0.98 cM).

DISCUSSION

The collaborative cross is a highly attractive proposition because a large set of recombinant inbred strains is a powerful resource for investigating the genetic basis of complex traits, making it possible to map genetic effects into small regions and investigate gene-by-gene and gene-by-environment interactions. One drawback to the proposal is that the cross might have to be extremely large to meet some of its objectives: initial suggestions were that 1000 strains would be required. Our simulation results for QTL mapping are more optimistic.

TABLE 7

Detection and mapping of an epistatic QTL in 1000 animals

	Power to detect QTL			Mean location error (cM)		
Strategy	0 RILs	500 RILs	1000 RILs	0 RILs	500 RILs	1000 RILs
cc[a]	—	0.29	0.54	—	1.72	1.34
cc04	—	0.25	0.47	—	1.34	1.07
cc08	—	0.18	0.40	—	0.98	0.82
cctop10	—	0.28	0.55	—	1.92	1.27
riail08	—	0.49	0.61	—	1.98	1.49
riail13	—	0.37	0.46	—	2.10	1.89
rihs05	—	0.11	0.09	—	3.01	2.25
rihs10	—	0.06	0.02	—	6.82	9.71
rihs30	—	0.00	0.01	—	—	2.13
BC	0.84	—	—	4.99	—	—
BC (SMA)	0.92	—	—	2.36	—	—
F_2	0.60	—	—	4.78	—	—
F_2 (SMA)	0.78	—	—	3.05	—	—

Tabulation of data in Figure 7 is shown. See Figure 7 legend and RESULTS for details.

[a] Additional simulations were performed for cc with results [written as number of RILs = power (location error)]: 100 RILs = 0.02 (8.16), 200 RILs = 0.05 (3.20), 250 RILs = 0.12 (3.07), 333 RILs = 0.19 (2.80).

We explored the power and mapping resolution of CC strategies for identifying QTL and compared the results to classical F_2 intercross and BC approaches, as well as to two other RI methods: one using RIs derived directly from HS animals and the other from an advanced intercross. We found a number of striking trends that together characterize the relationship between power, accuracy, and aspects of breeding design in two-way and eight-way crosses.

First, we show the increase in power and resolution that accompanies an increase in RI lines (Figure 5). We used a fixed population size of 1000 animals regardless of how many RI lines were included. This meant that we implicitly modeled the trade-off between replication and genetic diversity, since one is bought at the cost of the other (see BELKNAP 1998 for a theoretical treatment).

Second, there was a tendency for more highly recombinant strategies to detect the QTL less readily but, once detected, to map it more accurately. Higher recombination produces a finer-grain haplotype mosaic that, by its reduced linkage disequilibrium between neighboring markers, allows sharper localization. Most strategies achieve higher recombination through increased generation time (the exception being cctop10, which undergoes specific selection for recombination during breeding). More generations mean greater opportunity for genetic drift and hence more cases of some QTL allele frequencies dropping to undetectable levels (FALCONER and MACKAY 1996). This explains the close resemblance between the order of strategies in Table 6 and their relative power in Figures 5 and 7. Replication accentuates these undesirable effects by reducing the effective population size. Conversely, increasing the number of lines stabilizes the allele frequencies and results in a convergence of data series as seen in the power graphs of Figures 5 and 7.

Third, our simulations suggest that the greatest resolution is achieved through a hybrid collaborative cross strategy in which two maintenance years are inserted between mixing and inbreeding (cc08). In both simulations of single and epistatic QTL, this strategy delivered subcentimorgan mapping with 1000 RILs and only slightly less accurate predictions with 500 RILs. The standard cc localized with ~0.5 cM more error on average. However, the advantage of cc08 must be weighed against its lower power and the additional time and cost it would incur in stock creation. We acknowledge that the precision of our QTL localizations was limited by the density of markers in the GNF set. Notably this limit caused the apparent asymptote in Figure 6 (see RESULTS). We chose the GNF set because it was the best one available at the time of simulation. Since then, however, there have been and will continue to be denser marker sets available. It is possible that performing the simulations with more closely spaced markers may uncover a stronger advantage to using greater numbers of RILs than is suggested by Figure 5.

Our conclusions deserve certain qualifications, both general and specific in nature. On a general level, we chose a 5% QTL as a standard, but this may have provided a particularly stringent target, thereby making our results more conservative than necessary. Although a recent review surveyed the effect sizes of reported QTL phenotypes mapped using inbred crosses and found the mean was closer to 6% (FLINT *et al.* 2005), in fact a smaller effect size is likely to be a more typical value. That is partly because most mapping studies overestimate QTL effect sizes (due to their small sample size)

and partly because experiments that dissect QTL so often find that a single QTL is due to the combined effects of a number of physically linked QTL of small effect (Flint *et al.* 2005).

We also point out that our results do not tell us about the ability of the strategies to detect epistasis. Rather, our concern was to map a QTL in the presence of epistasis. In our simulations the extent to which the epistatic QTL remains detectable depends on the minor allele frequency of the second (masking) locus, which in turn depends on how well the chosen breeding strategy preserves heterogeneity. We may therefore underestimate the power and resolution attainable in practice when all chromosomes can be analyzed simultaneously.

A further general qualification is that our assessment of the trade-off between replication and genetic diversity, which affects the advantage of adding RI lines, is crucially dependent on the level of background genetic variation, which, in the case of mapping a 5% QTL, we have set at 45%. Replication squeezes out environmental noise. Phenotyping an individual once with environmental error is always going to be less accurate than phenotyping 10 identical individuals and taking their mean. That difference in accuracy is related to the size of the environmental component. For instance, if variation in the trait were entirely genetic, replication could confer no advantage. During mapping we take into account whether the population contains replicates. If not, we regress on single animals. But if there are replicates, we regress on strain means. Doing so amounts to deriving a pseudo-population of fewer animals with a QTL of greater effect size. How does this influence our results?

Background genetic variation, QTL effect size, and replication are related as follows. Let θ_Q be the effect size (*i.e.*, proportion of variance explained) in the mapping population of primary QTL, θ_B be the summed effect size of all background QTL, and θ_E be the remaining environmental variance, where $\theta_E = 1 - (\theta_Q + \theta_B)$, and suppose that the population of size N comprises N/n groups of n replicates. Then regressing on strain means is equivalent to mapping in a pseudo-population of N/n animals, where the effect size of the primary QTL is now

$$\theta_q = \frac{\theta_Q}{\theta_Q + \theta_B + (\theta_E/n)}.$$

For instance, consider a 5% QTL with no background QTL segregating in a population of 1000 genetically distinct animals (*i.e.*, $\theta_Q = 0.05$, $\theta_B = 0$, $N = 1000$, $n = 1$). Then consider the same QTL in a population that includes replicates: 1000 animals drawn in equal measures from 100 RILs (*i.e.*, $\theta_Q = 0.05$, $\theta_B = 0$, $N = 1000$, $n = 10$). Both instances describe a 5% QTL in a population of 1000 individuals. But by averaging phenotypes across genetically identical animals and thereby exploiting known population structure, the second instance is more usefully viewed as a 34% QTL in a pseudo-population of 100 "averaged" individuals.

How does inclusion of background QTL change this? In our simulations background QTL account for 45% of the phenotypic variance, such that the total heritability is 50% ($\theta_Q = 0.05$, $\theta_B = 0.45$). Without exploiting population structure, both instances still describe a 5% QTL segregating in 1000 individuals. However, treating the second instance as a pseudo-population of 100 now delivers an effect size of only 9%, showing a diminished advantage of replication in the face of a strong genetic background.

We chose a 50% heritable trait because it accorded with the estimated heritability of many physiological and behavioral traits. We found that power rose consistently with the number of lines used but that accuracy leveled off after 500 RILs, suggesting that 500 may be an economic optimum at which acceptable power is accompanied by fine localization. However, we note that if the trait were less heritable, this optimum would be lower; if the trait were more heritable, it would be higher.

It is noteworthy that the BC and F_2 apparently performed better than many of the RI strategies, particularly in the presence of epistasis and when the number of RI lines was low. The main advantage of the RI is the much better mapping resolution achievable with ~500 lines. However, the two-way strategies F_2, BC, and RIAIL owed some of their success to strain ascertainment bias. Selecting only founder pairs in which the QTL segregate conferred several advantages over the eight-way strategies. First, it ensured that QTL allele frequency in the founding population was always 50% (while it could be as low as 12.5% in an eight-way cross). Consequently, not only did the marker allele uniquely determine strain origin, but also the effect size of the QTL was often amplified. We set the allelic value (the amount that an allele adds to the phenotype) to be constant across designs and calibrated it so it would account for 5% of the variance in a population of the eight founders. Among the founders, most QTL had a minor allele frequency of <50%. However, in the two-way strategies, the QTL frequency rises to 50%, which increases the variance attributable to it and thus its effect size. Moreover, because some background QTL that segregate among eight founders will inevitably not segregate between two founders, the amount of background variation will be diminished. Note that this is somewhat offset by the fact that the same will be true of non-QTL markers, fewer of which will be polymorphic, leading to coarser-grain mapping.

These are important consequences of using simpler breeding programs. We could have set the effect size of the QTL according to the mapping population, but that would have required varying the allelic value across designs, an unfair depiction of biological reality. Nonetheless, there is one advantage of eight-way strategies that our simulations mask: namely that a QTL is more

likely to segregate among eight strains than among two. We have assumed that the QTL is known to segregate between the two chosen strains. Although pairs of strains are rarely chosen for a cross at random, it is a goal of the collaborative cross to make that decision less crucial to success of a mapping experiment. We discuss this point more fully in the APPENDIX and detail how it upwardly biases our assessment of power and accuracy for two-way strategies.

The F_2 and BC also have inherent disadvantages against RI strategies, two-way or otherwise. Because we simulate dominance effects, homozygous individuals lie at the extremes of the phenotypic spectrum. This means that the genetic variance of a QTL in an RI population will be greater than that in an F_2 population, which has heterozygotes, and particularly greater than that in a BC population, which lacks one homozygote class (BELKNAP 1998; FLINT *et al.* 2005). However, our inclusion of background QTL moderates these effects.

The poor performance of the RIHS strategies is also due partly to genetic drift, but the explanation is slightly more complex and relates to our inclusion of background QTL. The simulated RIHS design was faithful to the construction of the existing Northport HS. In particular, the number of mating pairs in the mixing stage of breeding was set at 50. In the case of rihs30, that introduced a dangerous bottleneck that allowed fixation or near fixation at many loci. In simulations mapping a foreground QTL that affects a monogenic quantitative trait we have found that increasing the number of mating pairs to 150 recovers some power and accuracy (data not shown). But in the presence of background QTL more extreme increases are necessary.

The rihs08 and rihs04 strategies echo these problems to a lesser extent. But the impact of the maintenance bottleneck in RIHS runs deeper. Table 3 lists 5% significance thresholds for different strategies. The thresholds for the rihs30 strategy are excessively high because the rihs30 population is quasi-inbred. Many markers on the genotyped chromosome (chromosome 1) are fixed or nearly fixed, and many of the background QTL (on chromosomes 2–7; see METHODS) have drifted far from their starting frequencies of 50% to have highly unequal effects. In our null simulations of rihs30, where the foreground QTL has zero effect, it is likely that some of the genotyped markers are in strong linkage LD with background QTL of amplified effect. Because these markers act as surrogates for a true QTL, they will produce good regression fits even in the absence of a foreground QTL and, as a result, cause threshold estimates based on that null distribution to soar.

Moreover, increasing the number of lines from 500 RILs to 1000 RILs will make matters worse. In the CC strategies, RILs are drawn from either a vast pool of genetically different individuals from an expansive mixing phase (as in cc and cctop10) or a large maintenance population. In the RIHS strategies, RILs are always drawn from a small maintenance population. Therefore, increasing the number of RILs releases little extra genetic variation but does increase the apparent sample size and therefore the significance of the observed association. Note that RIHS performed far worse here than in simulations reported previously (VALDAR *et al.* 2003). The reasons are twofold. First, in those earlier simulations we did not include background QTL and so avoided the pitfalls described above. Second, in those simulations we fixed the effect size of the QTL relative to the mapping population, which meant that QTL drifting to low frequency were given commensurately larger allelic values to produce a 5% effect. By contrast, our current simulations calibrate the allelic value in the founders and fix it thereafter such that the effect size of the QTL changes as its allele frequency drifts through breeding. Thus, our current simulations implicitly reward strategies that stem genetic drift and penalize those that encourage it.

In summary, we find that for QTL detection and high-resolution mapping, the CC would not greatly benefit from increasing the number of inbred strains above 500. In the presence of epistatic QTL, additional RIs certainly help, particularly with detection, but the gains are relatively modest. Given the cost and logistical difficulties inherent in creating, maintaining, and distributing a very large set of RIs, it is hard to justify the routine use of a CC >500 for QTL mapping, although there may be advantages using additional lines for other experimental purposes, such as expression analyses. The results presented here are not specific to mice and so will be of interest to geneticists designing panels of RILs for other model organisms and plants.

Finally, we found that no strategy described here will definitively identify a gene. The mapping resolutions are still insufficient to achieve that goal. In the best scenario that we achieve using an eight-way strategy, the location error is just under 0.6 cM (obtained by 2 additional years of strain mixing in the CC and using 1000 RIs). Note that this is not the same as a confidence interval. Rather, it means that if we were to map a QTL, the gene would lie on average 0.6 cM away from the highest association peak. Since the mean gene density in the mouse genome is $\sim$10 genes/Mb, ideally we need a method with a resolution of 100 kb (equivalent to a location error of 0.05 cM).

We thank Ken Paigen and Karl Broman for helpful discussions. This work was funded by a grant from The Wellcome Trust.

LITERATURE CITED

BELKNAP, J. K., 1998 Effect of within-strain sample size on QTL detection and mapping using recombinant inbred mouse strains. Behav. Genet. **28:** 29–38.

BOUCHER, W., and C. W. COTTERMAN, 1990 On the classification of regular systems of inbreeding. J. Math. Biol. **28:** 293–305.

CABALLERO, A., and M. A. TORO, 2000 Interrelations between effective population size and other pedigree tools for the management of conserved populations. Genet. Res. **75:** 331–343.

Churchill, G. A., D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie *et al.*, 2004 The collaborative cross, a community resource for the genetic analysis of complex traits. Nat. Genet. **36:** 1133–1137.

Coles, S., 2001 *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London.

Complex Trait Consortium, 2003 *A Collaborative Cross for High-Precision Complex Trait Analysis*, v. 31 (http://www.complextrait.org/).

Darvasi, A., and M. Soller, 1995 Advanced intercross lines, an experimental population for fine genetic mapping. Genetics **141:** 1199–1207.

Demarest, K., J. Koyner, J. McCaughran, Jr., L. Cipp and R. Hitzemann, 2001 Further characterization and high-resolution mapping of quantitative trait loci for ethanol-induced locomotor activity. Behav. Genet. **31:** 79–91.

Dudbridge, F., and B. P. Koeleman, 2004 Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am. J. Hum. Genet. **75:** 424–435.

Falconer, D., and T. F. Mackay, 1996 *Introduction to Quantitative Genetics*. Longman Group, New York.

Flint, J., W. Valdar and R. Mott, 2005 Experimental strategies for mapping and cloning quantitative trait genes in rodents. Nat. Rev. Genet. **6:** 271–286.

Kimura, M., and J. Crow, 1963 On the maximum avoidance of inbreeding. Genet. Res. **4:** 399–415.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

McClearn, G. E., J. R. Wilson and W. Meredith, 1970 The use of isogenic and heterogenic mouse stocks in behavioral research, pp. 3–22 in *Contributions to Behavior-Genetic Analysis: The Mouse as a Prototype*, edited by G. Lindzey and D. Thiessen. Appleton-Century-Crofts, New York.

Mott, R., C. J. Talbot, M. G. Turri, A. C. Collins and J. Flint, 2000 A method for fine mapping quantitative trait loci in outbred animal stocks. Proc. Natl. Acad. Sci. USA **97:** 12649–12654.

Pletcher, M. T., P. McClurg, S. Batalov, A. I. Su, S. W. Barnes *et al.*, 2004 Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. PLoS Biol. **2:** e393.

Stephenson, A., 2003 evd: extreme value distributions. R News **2:** 31–32.

Talbot, C. J., R. A. Radcliffe, J. Fullerton, R. Hitzemann, J. M. Wehner *et al.*, 2003 Fine scale mapping of a genetic locus for conditioned fear. Mamm. Genome **14:** 223–230.

Valdar, W. S., J. Flint and R. Mott, 2003 QTL fine-mapping with recombinant-inbred heterogeneous stocks and in vitro heterogeneous stocks. Mamm. Genome **14:** 830–838.

Wang, X., I. Le Roy, E. Nicodeme, R. Li, R. Wagner *et al.*, 2003 Using advanced intercross lines for high-resolution mapping of HDL cholesterol quantitative trait loci. Genome Res. **13:** 1654–1664.

Communicating editor: K. W. Broman

APPENDIX

The success of two-way strategies in our simulations is partly due to strain ascertainment bias. For these strategies, we selected founder pairs in which the QTL segregate, since to do otherwise would have meant simulating no QTL for most cases. However, this means that our assessment of power for these strategies is upward biased.

It is instructive to consider what would happen if, rather than ascertain strains, we choose the founder pair randomly from the eight. How would this affect power and resolution? The following reasoning assumes the QTL is one of the GNF SNPs and segregates among the eight founders. Power to detect describes a probability, $P(\text{detect})$. We can split this probability as follows:

$$P(\text{detect}) = P(\text{detect} \mid \text{segregates})P(\text{segregates}) + P(\text{detect} \mid \sim\text{segregates})P(\sim\text{segregates}),$$

where "segregates" means the QTL segregates and "∼segregates" means it does not. Clearly, $P(\sim\text{segregates}) = 1 - P(\text{segregates})$. Also the $P(\text{detect} \mid \sim\text{segregates})$, the probability to detect a QTL given that it is fixed, is equivalent to the significance threshold

$$P(\text{detect}) = P(\text{detect} \mid \text{segregates})P(\text{segregates}) + (0.05)(1 - P(\text{segregates})).$$

In the simulations performed here, we report only cases where the QTL segregates and so we report the power as $P(\text{detect} \mid \text{segregates})$. In the case of an eight-way strategy, $P(\text{segregates}) = 1$, and so $P(\text{detect} \mid \text{segregates}) = P(\text{detect})$. However, for a two-way strategy without strain ascertainment, $P(\text{segregate}) < 1$ and so $P(\text{detect}) \neq P(\text{detect} \mid \text{segregate})$.

We can work out $P(\text{segregrates})$ for a single SNP as follows. If the high allele exists in b of 8 strains, then the probability of choosing two strains but only one high allele is $P(\text{segregates} \mid b) = \frac{2}{7}b - \frac{1}{28}b^2$. Averaging this over all m GNF SNPs gives

$$P(\text{segregates}) = \sum_i^m P(\text{segregates} \mid b_i)P(i) = \frac{1}{m}\sum_i^m P(\text{segregates} \mid b_i) \approx 0.471,$$

where b_i is the value of b for marker i and $P(i)$ is the probability of choosing that marker. Therefore, for the two-way crosses, the power to detect without strain ascertainment is

$$P(\text{detect}) = P(\text{detect} \mid \text{segregates})(0.471) + (0.05)(0.529),$$

which would mean that an F_2, BC, or RIAIL whose power was 100% with strain ascertainment would have only ∼50% power without this advantage. This reduction is even more pronounced in the epistatic case. There, further strain selection to ensure segregation of the unlinked QTL results in $P(\text{segregates}) = 0.269$, which reduces an apparent 100% detection to ∼30%.

In summary, if we were to simulate the extreme opposite case of no strain ascertainment, for two-way strategies power would be about half that reported here when mapping a single QTL and less than a third in the case of epistasis.