File S2: Power simulation of QTL mapping

Suppose that there are n_R RIX crosses with n_p mice per cross such that $N = n_R * n_p$. In our simulations, we posit the following model at each locus, k:

$$Y^{(k)} = \beta_0 + X_m^{(k)}\beta_a + X_p^{(k)}\beta_a + Z\gamma + \epsilon$$

where:

Y corresponds to the $N \times 1$ phenotype vector to be mapped;

X_m is an $N \times 8$ matrix of allele probabilities, such that the ith row is the vector of probabilities whose elements correspond to the probabilities that the maternal allele of mouse i at locus k came from each of the eight CC founder strains (each row necessarily sums to one);

X_p is an analogous $N \times 8$ paternal allele probability matrix;

$Z'Z = K$, an $N \times N$ matrix that accounts for the genetic similarity between samples;

$\gamma \sim \mathrm{N}(0, \sigma^2_{\mathrm{poly}})$ is an $N_R \times 1$ vector, such that, when pre-multiplied by $Z_{N \times N_R}$, yields an $N \times 1$ vector of correlated polygenic random effects (one for each mouse);

$\epsilon \sim N(0, \sigma_e^2)$ is an $N \times 1$ error term;

β_a is an 8×1 vector corresponding to the effects of each of the eight founder alleles.

In our simulation, we start by selecting N_R RI strains, and generate RIX lines using a loop design. We calculated the genetic similarity for these RIX's using the genotyping on the MegaMUGA platform of the most recent common ancestors (MRCA's) for the RI strains. Then, for each replication, we randomly selected a location from among the approximately 74,000 autosomal loci on the array to be the causal location in the simulation replication. We assigned "true" maternal and paternal alleles for each RIX using the 8-state probabilies. Further, we randomly picked the subset of the 8 alleles to have the causal effect, based on a realistic distribution. In each simulation, the size of the causal allele effect was set to be a multiple of σ_{poly}, the standard deviation of γ. In all simulations, we employed $\sigma_e^2 = 1$ and $\sigma_{poly}^2 = 0.5$. We varied the effect of the causal allele as a multiple of σ_{poly}.

The model was fit using R/qtl2. For computational simplicity, rather than fitting the full model described above, we first averaged the phenotype within each strain. For each replication, we identified the LOD peaks that exceeded a significance threshold based on 1000 permutations (this threshold was based on a single replication, not re-calculated for each). For each replication, we identified whether the interval associated with a LOD peak overlapped the causal location. The power was calculated as the proportion of replications for which an LOD peak identified thus identified the causal location.

Our simulations showed that, even for very large effect sizes, the power to detect the causal allele was quite modest. For example, we visualize selected results in the figure below. These correspond to a loop of size 60 (i.e., 60 parental strains) with 6 mice per cross. This setting was chosen for visualization because it most closely resembles the sample size of our own data. The numbers on the x-axis indicate multiples of the σ_{poly}. As shown in *Figure S9*, even when the causal effect is 10 times σ_{poly} (an enormous effect), the estimated power to detect it is around 22%.