Which mouse multiparental population is right for your study? The Collaborative Cross inbred strains, their F1 hybrids, or the Diversity Outbred population

Gregory R. Keele*,[1]

[1]The Jackson Laboratory, 600 Main Street, 04609 Bar Harbor, ME, USA

*Corresponding author: greg.keele@jax.org

Abstract

Multiparental populations (MPPs) encompass more genetic diversity than traditional experimental crosses, enabling a wide array of experimental designs for deep genetic dissections of complex traits. For mouse models, two related MPPs have emerged: the Collaborative Cross (CC) inbred panel and the Diversity Outbred (DO) population. Additionally, the F1 intercrosses of the CC (CC-RIX) allow for studies of replicable outbred mice. Researchers often seek to characterize and genetically dissect traits through heritability estimation and mapping regulatory genetic loci. Here we evaluate the relative merits of these populations for these tasks through simulation, as well as provide recommendations for performing the quantative analyses. We find that sample populations that include replicate animals, as possible with the CC and CC-RIX, provide more efficient and precise estimates of heritability. CC-RIX populations with replicates enable heritability to be decomposed into additive and non-additive components. All populations of approximately 200 animals were well-powered to detect large effect loci that explain $\geq$ 40% of the phenotypic variation, but only large sample populations of 500 DO mice were well-powered to detect smaller effect loci ($\leq$10%) for highly polygenic traits. All results were produced with our R package musppr, which we developed to simulate data and evaluate genetic analyses from user-provided genotypes from these MPPs.

Keywords: CC; recombinant inbred intercross; CC-RIX; DO; MPP; multiparent advanced generation intercross; MAGIC; heritability; quantitative trait locus; QTL

Introduction

Multiparental populations (MPPs) are a powerful class of experimental cross for genetic studies of complex traits (de Koning and McIntyre 2017). Individuals are descended from multiple isogenic founder strains and thus possess recombinant genomes that encompass greater genetic diversity than traditional experimental crosses of two strains. MPPs have been developed across a wide range of model organisms, representing both animal models, such as heterogeneous stocks of mice and rats (Woods and Mott 2017), flies (Long *et al.* 2014), round worm (Noble *et al.* 2017), and yeast (Cubillos *et al.* 2013) as well as plants, including *Arabidopsis* (Kover *et al.* 2009), maize (Dell'Acqua *et al.* 2015), and rice (Bandillo *et al.* 2013). Here we focus on three related MPPs of the house mouse, *Mus musculus*: the Collaborative Cross (CC) panel of inbred strains (Collaborative Cross Consortium 2012; Srivastava *et al.* 2017), their F1 intercrosses (CC-RIX) (Threadgill *et al.* 2011; Schoenrock *et al.* 2018; Sun *et al.* 2021), and the Diversity Outbred (DO) population (Churchill *et al.* 2012).

The CC, CC-RIX, and DO are descended from the same eight inbred founder strains (short names in parantheses): A/J (AJ), C57BL/6J (B6), 129S1/SvImJ (129), NOD/ShiLtJ (NOD), NZO/HILtJ (NZO), CAST/EiJ (CAST), PWK/PhJ (PWK), and WSB/EiJ (WSB), which include both traditional laboratory and wild-derived strains, representing three subspecies of *Mus musculus* (Yang *et al.* 2007, 2011). As recombinant populations, the CC and DO are valuable resources for mapping quantitative trait loci (QTLs) and have been used to genetically dissect a wide range of phenotypes. Traits that have been studied in the CC include behavioral traits (Philip *et al.* 2011), hematology traits (Kelada *et al.* 2012), airway damage (Tovar *et al.* 2022) and allergies (Kelada *et al.* 2014), susceptibility to influenza (Ferris *et al.* 2013) and SARS-CoV (Gralinski *et al.* 2015; Schäfer *et al.* 2022), homeostatic immune regulation (Hampton *et al.* 2021), and drug response (Mosedale *et al.* 2017, 2019, 2021); in the DO, serum cholesterol (Svenson *et al.* 2012), insulin secretion (Keller *et al.* 2019), glutathione metabolism (Gould *et al.* 2021), response to benzene exposure (French *et al.* 2015), bone and skeletal traits (Al-Barghouthi *et al.* 2021), and working (Ouellette *et al.* 2020) and short-term (Hsiao *et al.* 2020) memory.

The number of realized fully-inbred CC strains was lower than planned (The Complex Trait Consortium 2004), with approximately 70 strains rather than 1,000 realized due to extinctions caused by allelic incompatibilities (Shorter *et al.* 2017). This reduction in the number of available unique genomes reduces the genetic mapping power possible with the CC; however, the inbred nature of the CC enables the use of replicates within and across experiments. Strain replicates can be used to characterize genetic architecture of traits and identify strain-specific phenotypes, due to variants that occurred and became fixed during the inbreeding generations or unique combinations of alleles. This approach has

Manuscript compiled: Friday 26th August, 2022

been used for a number of phenotypes and diseases, including spontaneous colitis (Rogala *et al.* 2014), peanut allergy (Orgel *et al.* 2019), immune cell diversity (Dupont *et al.* 2021), and susceptibility to virally-induced neurological phenotypes (Eldridge *et al.* 2020), tuberculosis, (Smith *et al.* 2019) and *Salmonella* (Zhang *et al.* 2019; Scoggin *et al.* 2022).

Genomics studies have been a growing area of research for both the CC and DO, leveraging their genetic diversity in the presence of large genetic effects on molecular, *i.e.*, "-omics" traits, such as gene expression (Aylor *et al.* 2011; Keller *et al.* 2018) and chromatin accessibility (Keele *et al.* 2020), proteins (Chick *et al.* 2016) and their phosphorylation sites (Zhang *et al.* 2022), and lipids (Linke *et al.* 2020), across a range of tissues and organs. Further studies have investigated how age influences gene expression and protein abundance in kidney and heart tissues from DO mice (Takemon *et al.* 2021; Gerdes Gyuricza *et al.* 2022). Genetic relationships between molecular traits have been characterized in embryonic stem cells derived from DO mice (Skelly *et al.* 2020; Aydin *et al.* 2022). Genetic effects are strongly consistent between populations, supporting the use of both populations for joint analysis and validation of findings (Keele *et al.* 2021).

QTL mapping power in the CC was initially evaluated prior to the development of the final panel of fully inbred strains, using far greater numbers of simulated genomes than were realized (Valdar *et al.* 2006), which we updated simulating from readily available Collaborative Cross strains (Keele *et al.* 2019). Mapping power has also been assessed in the DO population (Gatti *et al.* 2014). Here we extend our previous approach of simulating from observed genomes to evaluate and compare the performance of commonly used genetic analyses, heritability estimation and QTL mapping, among the CC, CC-RIX, and DO. This work will aid researchers in designing and tailoring their experiments to maximize their efficiency for various goals (*e.g.*, characterizing genetic architecture or mapping causal genetic variation), as well as provide recommendations for best practices in performing the quantitative and statistical analyses.

Methods

Sample populations

Artificially generating genetic data for MPPs from scratch is challenging on many levels. An ideal approach would involve full simulation of the various breeding designs from initial founder mice all the way to the reconstruction of founder haplotypes from marker genotypes in the final recombinant mice. To avoid this complex and computationally intensive approach, we simulated data for CC, CC-RIX, and DO (Figure 1) from observed populations of CC and DO. This approach has the additional advantage of capturing effects of any deviations from the breeding designs that occurred in practice, potentially better approximating real animals available to researchers.

The CC sample consisted of 116 mice (Keele *et al.* 2021) (female/male pairs from 58 strains) that were genotyped on an 11,000 marker array (MiniMUGA) (Sigmon *et al.* 2020). We used two DO sample populations, including 192 mice (Chick *et al.* 2016) genotyped on a 57,000 marker array (MegaMUGA) (Morgan and Welsh 2015) and 500 mice (Keller *et al.* 2018) genotyped on a 143,000 marker array (GigaMUGA). Founder haplotypes were probabilistically inferred using a hidden Markov model (HMM) implemented in the qtl2 R package (Broman *et al.* 2019). Genetic mapping in experimental crosses is commonly performed in terms of founder haplotypes rather than specific genetic variants, with uncertainty accounted for using a mixture model (Lander and Botstein 1989) (*i.e.*, interval or linkage mapping) or regression-based approximations (Haley and Knott 1992; Martínez and Curnow 1992). Because the CC were genotyped on a sparser array than the DO, their founder haplotype probabilities possess greater uncertainty. To make the populations as comparable as possible and avoid our assessment of heritability estimation and QTL mapping reflecting variation in founder haplotype uncertainty, we imputed founder haplotype probabilities at the same 64,000 loci (*i.e.*, pseudomarkers) spanning the genome and then imputed founder haplotype pairs (*i.e.*, diplotypes) identities based on having the greatest probability. We derived the founder diplotypes for the 1,653 CC-RIX F1s, representing all possible pairings between 58 CC strains (ignoring parent-of-origin features, *e.g.*, sex chromosomes and mitochondria). See Appendix A for greater detail on how diplotypes were processed for all sample populations.

Heritability

Heritability, the proportion of variation in a population explained by genetic relationship, is commonly used to assess evidence that a phenotype is genetically controlled. It is important to note that heritability is specific to the population, and thus cannot be extrapolated to other populations. Furthermore, it can be influenced by other sources of variation in the data, such as the measurement error of the phenotype, which will likely reduce heritability estimates.

Depending on the sample population, heritability can be estimated using different approaches and even decomposed into multiple components, such as the proportion of variation explained by all genetic effects, *i.e.*, broad-sense heritability (commonly symbolized as H^2) and the proportion of variation explained by additive genetic effects, *i.e.*, narrow-sense heritability (h^2) (Lynch and Walsh 1998). In this work, we use the h^2 notation more generally and will include subscripts to specify components of heritability in certain contexts. Parent-offspring regression (Falconer and Mackay 1996) is one approach to estimating heritability that is generally not possible for studies of CC, CC-RIX, or DO because parent-offspring pairs are often not observed and measured within the same experiment. For studies of inbred strains with replicates, as is possible with the CC, an intraclass correlation can be used (as in Yam *et al.* (2021)), though it is less appropriate in the outbred CC-RIX and not possible in the DO.

Heritability model: Linear mixed effects models (LMMs) offer an appealing general approach that can be adjusted for each of the CC, CC-RIX, and DO. We estimate heritability using the following general LMM:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{u} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y} = \{y_i\}_{i=1}^N$ is the phenotype vector, μ is the intercept, and $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ are N-length random vectors. $\mathbf{u}$ can be referred to as the polygenic effect, here representing structured error (*i.e.*, population structure) and is modeled as $\mathcal{N}(\mathbf{0}, \mathbf{K}\tau^2)$ where $\mathbf{K}$ is an $N \times N$ (often additive) genetic relationship matrix (*i.e.*, kinship matrix) and τ^2 is the corresponding variance component. $\boldsymbol{\varepsilon}$ is unstructured error, distributed according to $\mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2)$ where $\mathbf{I}$ is an $N \times N$ identity matrix and σ^2 is its variance component. Heritability is then calculated as the ratio of variation due to genetics to total variation:

$$h^2 = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{2}$$

To obtain unbiased estimates of heritability, the variance component parameters are estimated through optimizing the restricted maximum likelihood (REML) (Patterson and Thompson 1971). The

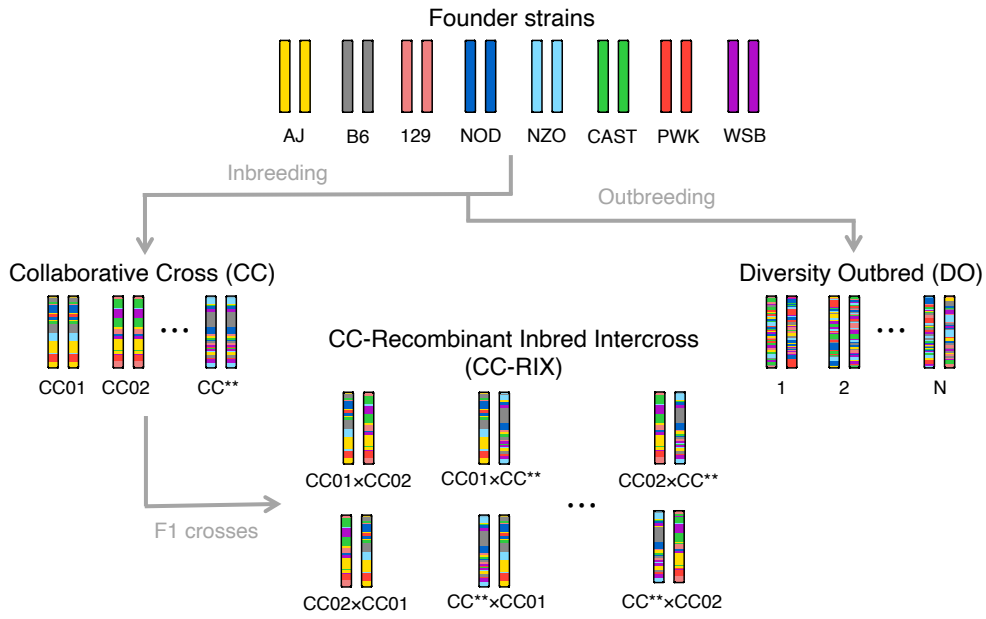

Figure 1 Cartoon diagram relating the CC, CC-RIX, and DO. All the recombinant populations are descended from the same eight inbred founder strains. Each genetically unique individual or strain is depicted as a single pair of cartoon chromosomes. The genomes of the recombinant populations are mosaics of the founder genomes. For the CC-RIX, reciprocal F1s are depicted, *e.g.*, CC01×CC02 versus CC02×CC01. CC** represents the final CC strain. N is the total number of DO mice.

heritability LMM can be fit with various software packages; here we compare results from the qtl2 R package (Broman *et al.* 2019), our miQTL R package (available at https://github.com/gkeele/miqtl), and the sommer R package (Covarrubias-Pazaran 2016).

For sample populations that include strain or F1 replicates, such as in the CC and CC-RIX, the kinship matrix can be characterized as $\mathbf{K} = \mathbf{Z}\mathbf{K}_M\mathbf{Z}^{\mathrm{T}}$ where $\mathbf{Z}$ is the $N \times M$ strain/F1 identity matrix mapping the N individuals to M strains/F1s ($M < N$) and $\mathbf{K}_M$ is the $M \times M$ kinship matrix encoding the overall genetic relationship between the M strains/F1s.

If replicates are included in the sample population, Eq 1 can be expanded to include two components of heritability:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{u}_{\mathrm{add}} + \mathbf{u}_{\mathrm{rep}} + \boldsymbol{\varepsilon} \tag{3}$$

where $\mathbf{u}_{\mathrm{add}}$ is equivalent to $\mathbf{u}$ from Eq 1 in the presence of replicates—$\mathbf{u}_{\mathrm{add}} \sim \mathcal{N}(\mathbf{0}, \mathbf{Z}\mathbf{K}_M\mathbf{Z}^{\mathrm{T}}\tau^2_{\mathrm{add}})$—and $\mathbf{u}_{\mathrm{rep}}$ is a random effect specific to strain/F1—$\mathbf{u}_{\mathrm{rep}} \sim \mathcal{N}(\mathbf{0}, \mathbf{Z}\mathbf{Z}^{\mathrm{T}}\tau^2_{\mathrm{rep}})$. The two components of heritability are then estimated as

$$h^2_{\mathrm{add}} = \frac{\tau^2_{\mathrm{add}}}{\tau^2_{\mathrm{add}} + \tau^2_{\mathrm{rep}} + \sigma^2} \tag{4}$$

$$h^2_{\mathrm{rep}} = \frac{\tau^2_{\mathrm{rep}}}{\tau^2_{\mathrm{add}} + \tau^2_{\mathrm{rep}} + \sigma^2} \tag{5}$$

where h^2_{add} is the proportion of variation explained by additive genetic effects, *i.e.*, narrow-sense heritability, and h^2_{rep} is the proportion of variation explained by strain/F1 identity, such as due to epistasis between loci with combinations of alleles unique to a given strain/F1. From the three software packages we used, only sommer allows for flexible specification of multiple random effects, each with their own covariance matrix.

Kinship matrix for heritability: The true kinship matrix $\mathbf{K}$ is unknown and must be estimated from genotypes (haplotypes or SNPs), pedigree information, or both (Cheng *et al.* 2013). Here our focus is not a thorough analysis of how to best estimate the kinship matrix, though we do evaluate a number of options. For a SNP-based kinship matrix, we use the form described in Endelman and Jannink (2012) with selection parameter set to 0. See Appendix A for details on imputing SNPs from the founder haplotypes. For haplotype-based kinship matrix, we estimated one using qtl2 (Broman *et al.* 2019), which is calculated as $\widehat{\mathbf{K}}_{\mathrm{hap}} = \sum_{p=1}^{P} \mathbf{A}_p\mathbf{A}_p^{\mathrm{T}}$, where $\mathbf{A}_p$ is the scaled founder haplotype dosage matrix at locus p. We also evaluated the average semivariance (ASV) form (Feldmann *et al.* 2022) of each kinship matrix: $\widehat{\mathbf{K}}_{\mathrm{ASV}} = \frac{N-1}{\mathrm{tr}(\widehat{\mathbf{K}})}\mathbf{P}\widehat{\mathbf{K}}\mathbf{P}^{\mathrm{T}}$, where $\mathbf{P} = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\mathrm{T}}$ is the mean-centering matrix.

Heritability simulations: We adapted our approach used in Keele *et al.* (2019) to simulate phenotype data with specified heritability; see Appendix B for greater detail. When simulating a single additive component of heritability for a given population, we simulated 1,000 data sets for each specified level of heritability. Heritability was evaluated across a grid, ranging from 0 to 1 and separated by regularly-spaced intervals of 0.05. Heritability was then estimated using one of qtl2, miQTL, or sommer, and summarized with the mean and the 95% estimate interval (defined by the 0.025 and 0.975 empirical quantiles). Within this simulation framework, we also compared using kinship matrices estimated from SNPs and founder haplotypes, as well as ASV forms of both.

We extended the previous simulation approach for heritability with two components by fixing the total heritability at a specified level (*e.g.*, 90%) and varied the ratio $h^2_{\mathrm{add}}/h^2_{\mathrm{rep}}$ across a grid, ranging from 0 to 1 and separated by regularly-spaced intervals of 0.1. Only 100 data sets were simulated per parameter setting because fitting the two component heritability model in sommer is more computationally intensive and thus slower to fit. Means and 95%

estimate intervals were used as summaries as before, but now for each component of heritability separately as well as their sum.

We performed various comparisons of heritability estimation across the sample populations of CC, CC-RIX, and DO by generating simulated populations derived from their genomes. For example, we randomly sampled 116 DO mice from the sample population of 192 to directly compare to the 116 CC mice (from 58 strains). When evaluating two components of heritability, we simulated large populations of 522 mice given that each genome must be observed with multiple replicates to distinguish the strain- or F1-specific component from the additive one. For the CC-RIX, given that it is unlikely researchers would collect a full data set of all F1s (1,653 for 58 CC strains), we evaluated three different simulated populations. First a balanced set where each CC strain is represented by two F1s as could be defined with a rotational breeding scheme (*e.g.*, CC001×CC058, CC001×CC002, CC002×CC003, ..., CC057×CC058). Each F1 can be observed through multiple replicates. A key feature of a balanced sample of CC-RIX F1s is that they possess the same overall allele frequencies as a corresponding sample of the parental CC strains, allowing us to evaluate how the CC-RIX population structure and heterozygosity affects heritability estimation and mapping power. We also evaluated simulated populations composed of randomly sampled CC-RIX F1s, either in combination with the parental CC strains or by themselves.

QTL mapping

The chromosomes of individuals from MPPs are mosaics of the founder haplotypes formed through recombination events that occurred throughout breeding generations (Figure 1). This mixing process randomizes loci from each other (ignoring linkage disequilibrium (LD) and population structure), allowing genetic variation at loci to be causally associated with traits. We evaluate mapping power in the context of conventional single locus genetic mapping, though note that these approaches could be extended to multi-locus models and epistasis.

QTL model: The underlying model used during QTL mapping is similar to Eq 1, however a locus effect is evaluated at loci spanning the genome (*i.e.*, a genome scan):

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{QTL}_{[p]} + \mathbf{u}_{[c[p]]} + \boldsymbol{\varepsilon} \tag{6}$$

where $\mathbf{QTL}_{[p]} = \mathbf{A}_p \boldsymbol{\beta}_{\text{QTL}}$ is the effect of locus p, $\mathbf{A}_p$ is the $N \times 8$ scaled founder haplotype dosage matrix at locus p (assuming an additive model), $\boldsymbol{\beta}_{\text{QTL}}$ is the 8-vector of founder haplotype effects, $\mathbf{u}_{[c[p]]}$ is the random polygenic effect vector with respect to chromosome c (which contains locus p), and all other terms as previously defined. We used the leave-one-chromosome-out (LOCO) approach (Yang *et al.* 2014; Gatti *et al.* 2014) in which $\mathbf{u}_{[c[p]]} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{[-c[p]]})$, meaning $\mathbf{K}$ is estimated from all markers excluding those from locus p's chromosome, which avoids any of the effect of locus p from being absorbed into the random terms and increases mapping power. We used qtl2 to perform all genome scans.

QTL significance thresholds: Instead of using permutations (Churchill and Doerge 1994) to estimate significance thresholds that control the family-wise error rate (FWER), we used parametric bootstrap samples from the null model (Eq 6 excluding the QTL term). We note that bootstrap samples are generated given the true value of heritability, which would normally need to be estimated from the data and thus subject to error. Our goal in this context is to produce thresholds that generalize across simulated data from a given population and can thus be compared across the populations. Thresholds were calculated as quantiles from extreme value distributions fit from the maximum LOD scores across the null bootstrap samples (Dudbridge and Koeleman 2004).

Confidence interval for QTL location: The goal of QTL mapping is often to identify candidate genes or functional variants that underlie a QTL. But just as estimation of the effects of the QTL are subject to error, so is the estimation of QTL location. Furthermore, most often the causal genetic variant itself is not being tested as a QTL, but rather a locus that is in LD with it. This is particularly relevant in MPPs, where it is conventional to initially detect QTL using sparser scans based on founder haplotypes at intervals across the genome (Eq 6). A confidence or support interval for the location of QTL statistically summarizes this uncertainty and can prioritize a genomic region for specific candidate genes or genetic variants. Likelihood-based support intervals are often used and have been statistically characterized for F2 intercrosses and backcrosses (Dupuis and Siegmund 1999), though not for MPPs such as the CC, CC-RIX, or DO. Empirical sampling-based approaches have also been proposed, such as non-parametric bootstrapping (Visscher *et al.* 1996), though it performed poorly with sparse markers (Manichaikul *et al.* 2006). Here we evaluate approaches for estimating QTL intervals in these populations, including likelihood-based—LOD support intervals and approximate Bayesian credible intervals (BCIs) (Broman and Sen 2009)—and sampling-based intervals—parametric bootstraps, parametric permutations, and Bayesian bootstraps (Rubin 1981). See Appendix C for more details on each method.

QTL simulations: We extended our approach to simulating mapping data in the CC (Keele *et al.* 2019) to the CC-RIX and DO populations. Our focus is not a deep characterization of mapping power for a specific population, but rather a broader comparison across populations of quantities related to mapping performance, including power, false discovery rate (FDR), mapping resolution, and interval summaries for QTL location. To contrast the populations, we primarily considered two scenarios: 1) a large effect QTL that explains 40% of the phenotypic variation (*i.e.*, 40% effect size) with some background genetic architecture (polygenic effect explains 30% of phenotypic variation) in 174 mice, and 2) a small effect size QTL (10%) for a trait with highly polygenic architecture (80%) in 500 mice (and in a sample of 174 DO mice for comparison between scenarios). We wanted to evaluate large QTLs in samples of 200 or fewer mice, though we note that 174 is somewhat arbitrary. We had genetic data for 58 CC strains and a smaller DO sample of 192 mice. Using three replicates per CC strain resulted in 174 mice, whereas four replicates (232 mice) would have exceeded the smaller DO sample.

For each scenario, we randomly selected 1,000 loci which were used to simulate QTLs in the five previously described populations of CC, CC-RIX, and DO. Each QTL was simulated with a bi-allelic effect evenly split across the founder strains. The variation explained by the QTL and background genetic architecture are strictly controlled with respect to the sample population. We also calibrated effect sizes with respect to a balanced homozygous population to investigate how unbalanced allele frequencies and heterozygosity may affect QTL mapping power. See Appendix D for further explanation of our simulation approach for QTL data.

Results and discussion

We simulated data using observed genomes of CC, CC-RIX, and DO mice to evaluate heritability estimation and QTL mapping; see

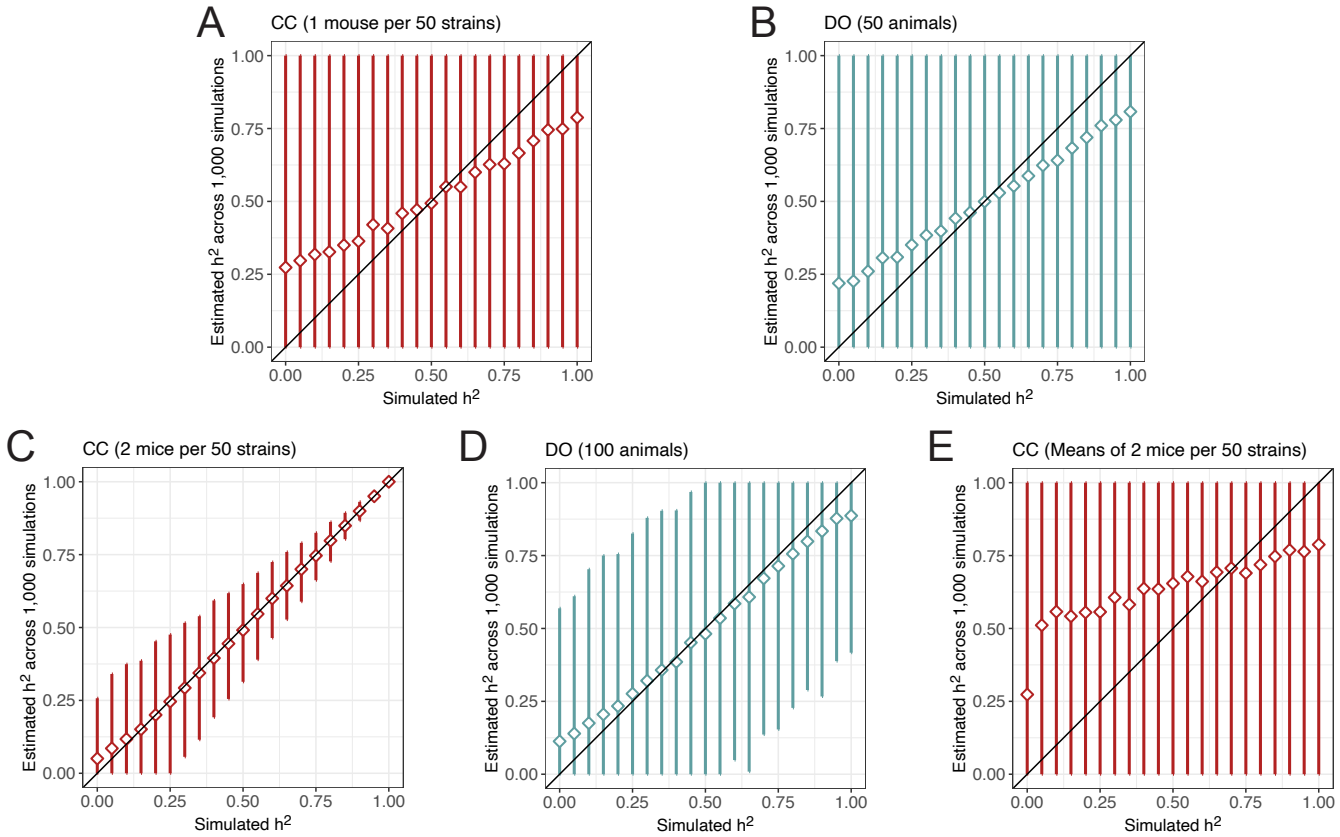

Figure 2 Performance of heritability estimation in data simulated for (A) one mouse per 50 CC strains, (B) 50 DO mice, (C) two mice per 50 CC strains, (D) 100 DO mice, and (E) strain means from two mice per 50 CC strains. Diamonds represent the mean estimated heritability across 1,000 simulations from the true heritability. Vertical line segments represent middle 95% intervals across the 1,000 simulations. Black lines indicating the heritability estimate is equal to the true value (y = x) included for reference.

Methods for more details on the populations, the simulations, and statistical methods used.

We first compared heritability estimation across kinship matrices (haplotype-based and SNP-based, both non-ASV and ASV forms) and statistical software (qtl2, miQTL, and sommer) that can fit the same LMM (Figure S1). Haplotype-based and SNP-based kinship matrices corresponding to the same ASV status performed similarly, which is not surprising given that the SNPs were imputed from the haplotypes, but does confirm that these estimators produce similar genetic relationship matrices. In terms of software packages, miQTL and sommer produce essentially identical estimates of heritability, performing best with ASV kinship matrices. In contrast, qtl2 performs best in the DO with non-ASV haplotype-based matrices and is negatively biased when estimating heritability in the CC. For these reasons, we present the remaining heritability results using miQTL, which is computationally faster than sommer, and using ASV haplotype-based kinship matrices.

Replicates improve precision of heritability estimates

We first evaluated heritability estimation in small samples of 50 and 100 mice in the CC and DO (Figure 2). For a sample of 50 mice, the mean heritability across the simulations was more biased at the tails of the distribution. More importantly, the 95% estimate intervals cover the full support of heritability; in other words, regardless of the true value of heritability, the interval ranges from 0% to 100% (Figure 2A-B). Doubling the number of animals to 100 increases the precision of heritability estimation (Figure 2C-D), most notably in the CC population with two replicates per strain. For a 50% heritable trait in 100 DO mice, the 95% estimate interval still ranges from 0% to 100%, whereas in the CC, it ranges from 30% to 65%. In the CC, there was also greater variability in heritability estimates when the true heritability was lower, though the widest intervals are still notably narrower than in the DO sample of the same number of mice. We also evaluated heritability estimation using strain means (Figure 2E), which are commonly used for QTL mapping, and observed upward bias and intervals that cover the full support of heritability. The upward bias is consistent with the expected reduction in noise on a mean, and the wide intervals are due to the loss of the within-strain correlation information. This finding suggests that replicates should not be reduced to strain summaries for the purpose of estimating heritability.

We expanded our evaluation of heritability estimation to include three forms of CC-RIX populations while fixing the number of total mice at 174 (some populations with three replicates per 58 strains/F1s; Figure 3) and 500 (some populations with ten replicates per 50 strains/F1s; Figure S2). Compared to populations of 50 or 100 mice, larger populations of 174 or 500 are able to estimate heritability more accurately and precisely. To more directly compare how the number of replicates versus the total number of mice influence the accuracy and precision of heritability estimation across the populations, we performed 1,000 simulations for varying numbers of replicates and total mice at three levels of heritability (25%, 50%, and 75%) (Figure 4). Notably, replicates can improve the precision of heritability more than total number of

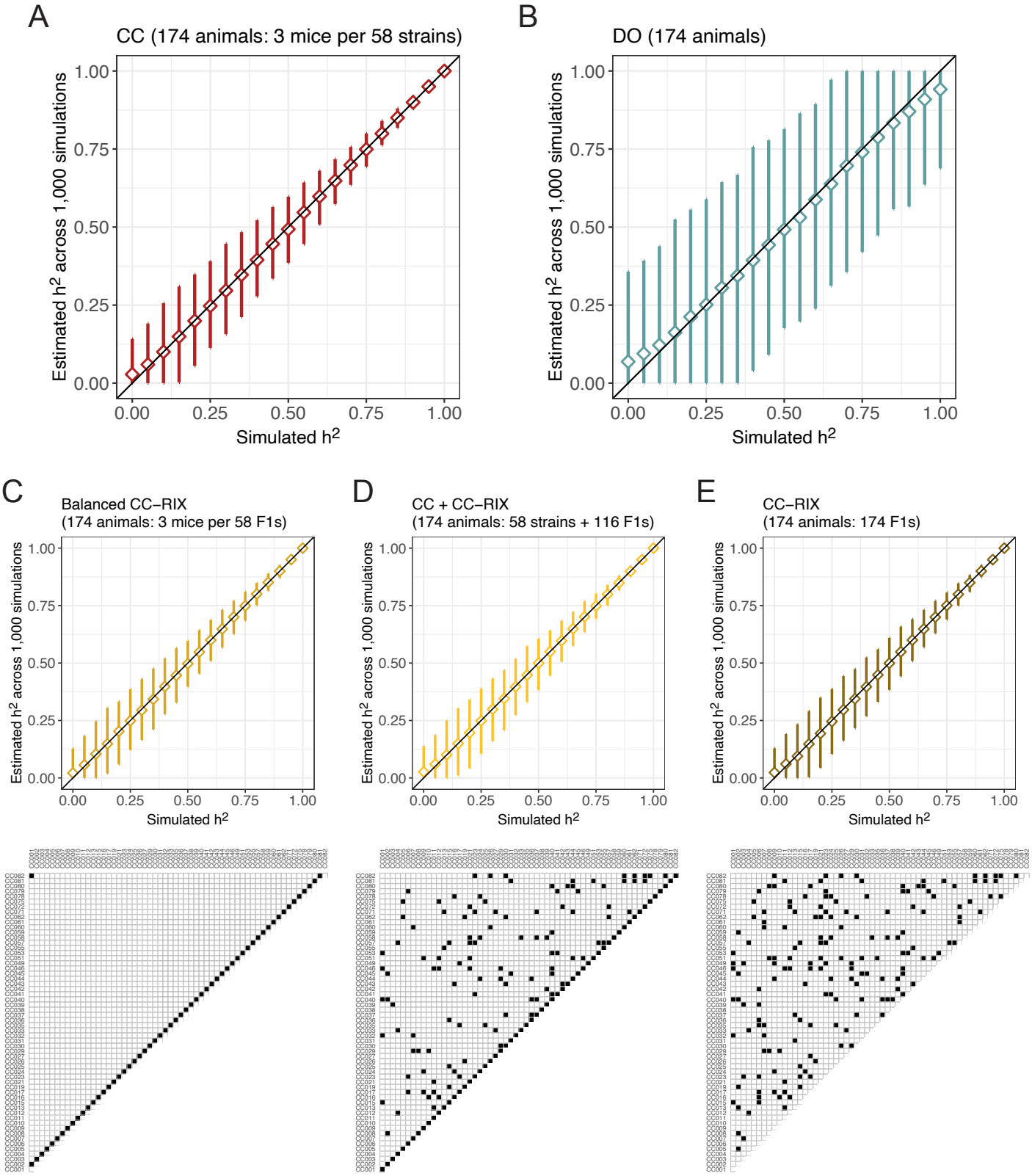

Figure 3 Performance of heritability estimation in data simulated for 174 mice from (A) CC, (B) DO, and (C-E) CC-RIX populations. Diamonds represent the mean estimated heritability across 1,000 simulations from a given true heritability. Vertical line segments represent middle 95% intervals across the 1,000 simulations. Black lines indicating the heritability estimate is equal to the true value (y = x) included for reference. Each CC-RIX population (C-E) is represented with a diallel grid of the CC strains and their F1s (ignoring effects of sex chromosomes or parent-of-origin), with observed strains and F1s as black cells (bottom row). See Figure S2 for results from simulations of 500 mice.

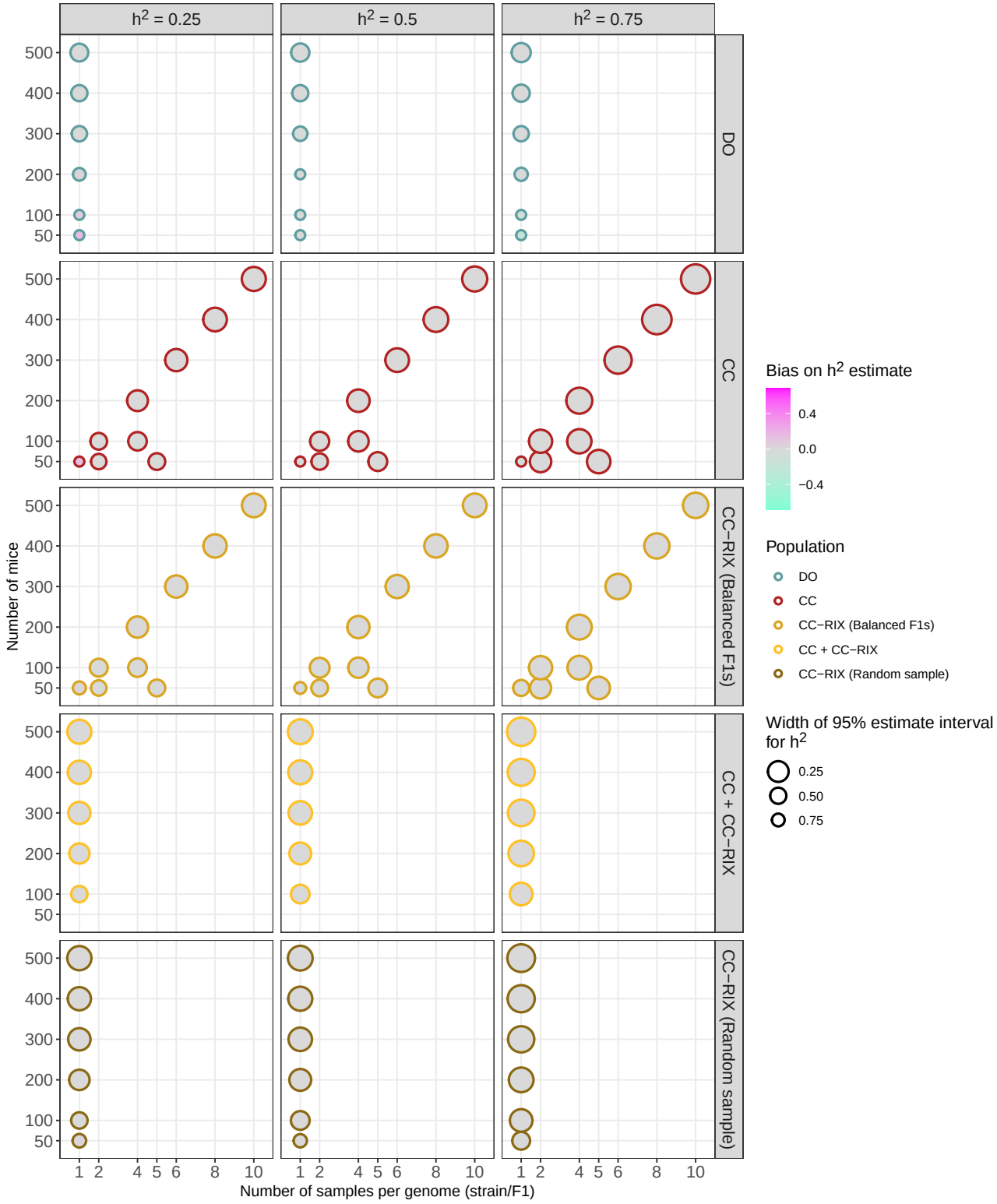

Figure 4 Comparing bias and variance in heritability estimation across CC, CC-RIX, and DO populations, number of replicates, and total sample size. Points represent summaries over 1,000 simulations for a given population type (rows), true heritability (columns), number of replicates per genome (x-axis), and total number of mice (y-axis). Point centers are colored by the bias in the mean heritability estimate ($\frac{1}{1000}\sum_{i=1}^{1000}\widehat{h_i^2} - h^2$), with upward bias being magenta, downward bias being aquamarine, and low bias being gray. The size of the point are inversely proportional to the width of the middle 95% estimate interval, with larger points being more precise.

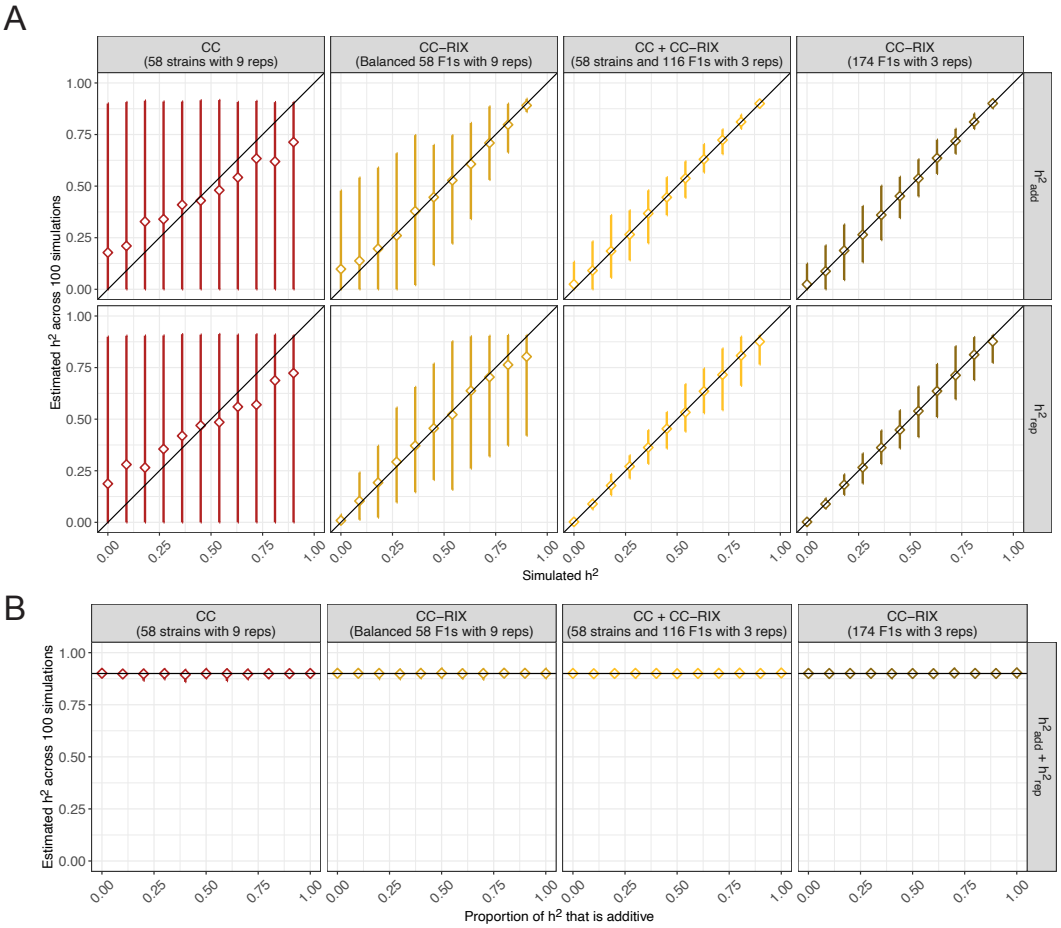

Figure 5 Performance of two component heritability estimation in data simulated for 522 mice from CC and CC-RIX populations with strain/F1 replicates. The sum of both components was fixed at 90%. Heritability was estimated either in terms of (A) each component, with the additive component (h^2_{add}) shown in the top row and the strain- or F1-specific (h^2_{rep}) shown in the bottom row, or as (B) the sum total. Diamonds represent the mean estimated heritability across 100 simulations from the true heritability. Vertical line segments represent middle 95% intervals across the 100 simulations. Diagonal black lines indicating the heritability component estimate is equal to the true value (y = x) included for reference. Horizontal black lines indicating the true value of the sum of heritability components (y = 0.9) included for reference.

mice—a sample of five replicates per 10 CC strains (50 mice total) has similar or better precision than DO populations of 400-500 mice. For example, the 95% estimate interval width in a sample of 5 mice from 10 CC strains is 15.1% compared to 100.0% in 100 DO mice, 43.3% in 400 DO mice, and 32.7% in 500 DO mice when the true heritability is 75%. Overall, bias was not an issue for any of the populations except for small samples of 50 DO mice.

The CC and CC-RIX populations provide more precise heritability estimation than the DO because they possess greater population structure (Figure S3) in the form of genetic replicates and/or shared parental CC strains for the CC-RIX. In data simulated for 174 mice and heritability set to 80%, all populations produce low bias, but the 95% estimate interval width for the DO is greater than 45% and around 10% in the CC and CC-RIX (Figure S3B). Notably, population structure also has implications for QTL mapping power. For example, not accounting for population structure in QTL mapping, produces many false positives, including many with LOD score > 10, in the unbalanced CC-RIX populations (Figure S3C), emphasizing the importance of accounting for population structure though the random effect in the LMM. We will discuss QTL mapping power in greater detail further on.

Replicates in the CC-RIX distinguish additive and non-additive components of heritability

We next performed simulations with two components of heritability: an additive component as before that reflects the additive kinship matrix, as well as a strain- or F1-specific component. Because replicate DO samples are not observable, we only evaluated CC and CC-RIX populations (Figure 5). All simulations had a cumulative heritability of 90% in populations of 522 mice. Across 100 simulations in the CC, the estimated components are generally unbiased but their 95% intervals range from 0% to 90%. The cumulative heritability was very accurately and precisely estimated at 90% (Figure 5B), suggesting that the individual components are generally indistinguishable in the CC. This is unsurprising given that the strains of the CC are largely equally related and thus the majority of population structure in a sample population comes from replicates, confounding the additive and strain-specific effects. This is further confirmed by how fitting an additive-only heritability model to data simulated with both additive and strain components generally returns the cumulative heritability total rather than the additive component (Figure S4A).

All the CC-RIX populations are better able to distinguish the

additive and non-additive components of heritability than the CC, particularly the less balanced populations that possess greater population structure in the form of F1s that share CC parental strains and fewer replicates per F1. Fitting an additive-only heritability model in simulated CC-RIX data with two components is also better able to capture only the additive component, though some of the non-additive variation does appear to upwardly bias the estimate (Figure S4B).

All populations of 174 mice are well-powered to detect large effect QTLs

We evaluated QTL mapping power in simulated traits with simple (40% QTL and 30% polygene) and complex (10% QTL and 80% polygene) genetic architectures. Our goal was to cover the spectrum of genetic complexity ranging from traits with simple genetic control though not necessarily Mendelian (*e.g.*, gene expression) to complex traits with high heritability driven by many loci. The large effect QTL results are based on simulations of 174 mice (three replicates per strain/F1 for the CC and the balanced CC-RIX) and the small effect QTL results are primarily based on 500 mice (ten replicates in CC and balanced CC-RIX).

For large effect QTLs, we find that all populations are well-powered (Figure 6A). With a lenient threshold of LOD score > 6, justifiable when detecting local genetic effects on molecular traits like gene expression, mapping power is essentially perfect in all populations, though at the expense of high false positive rates (FDR near 50% or greater). Using stringent thresholds that maintain FWER < 0.05 (*i.e.*, 95% significance thresholds), power ranges from 100% in the DO to 82% in the balanced CC-RIX, and notably, false positives are well-controlled. To assess the mapping resolution of the populations, we used the 1.5 LOD support intervals for QTL location. Unsurprisingly, the DO had narrower LOD support intervals, followed by the CC and CC-RIX. Here we use the LOD support interval as a summary of mapping resolution, largely for its ease in estimation. Further on we more deeply evaluate the statistical performance of QTL location interval estimates.

We also used the large effect QTL scenario to evaluate the power and false discovery properties of QTL mapping with CC strain means or individual-level data. In the CC, it is common to perform QTL mapping based on strain-level summaries (*e.g.*, means) rather than individual mouse-level data, which has the added benefit of reducing the data and thus the computational burden. Based on simulations of 1,000 QTLs in 116 mice (two animals per strain), we observed similar mapping power (for FWER < 0.05, 87% power in CC strain means compared to 91% in CC individuals; Figure S5). Though this evaluation is not exhaustive (the possible settings becomes complex—varying QTL and polygenic effect sizes as well as the number of replicates), we find that both strain-level and individual-level QTL mapping are valid approaches.

Only large sample populations of DO (500 mice) are well-powered to detect small effect QTLs

For small QTLs in highly heritable genetic backgrounds (10% QTL and 80% polygene), a sample population of 500 DO mice is well-powered (97%), whereas CC and CC-RIX populations essentially have no power when strictly controlling false positives (Figure 6B). Using lenient thresholds of LOD score > 6 increases power in the CC and CC-RIX (to roughly 10-20%) but false positives become highly common (around 90% FDR or greater for all populations). To more directly compare to the prior large QTL scenario (40% QTL and 30% polygene), we also evaluated the small QTL settings in 174 DO, which resulted in mapping power of 17%. This highlights the importance of large sample sizes when mapping QTL with effect sizes of 10% or lower. It also emphasizes that non-replicate genomes are needed for power when mapping in the presence of strong population structure—174 DO mice are better powered to detect a 10% QTL for highly heritable traits than 500 CC or CC-RIX mice. A key takeaway from these results is the contrasting value of genetic replicates for heritability estimation and QTL mapping—greatly increasing precision of heritability but contributing less to mapping power than a more independent, genetically unique sample would.

Likelihood-based interval estimates of QTL location outperform sampling-based ones

Interval estimates for QTL location are important when trying to survey and prioritize candidate genetic variants or genes near a QTL. Defining a conventional statistical confidence interval for QTL location is irregular because the likelihood is non-smooth based on the discrete positions of markers. Approximations to a confidence interval include likelihood-based methods that estimate intervals using the LOD scores from markers around a detected QTL as a profile likelihood for QTL location. Here we evaluate two such methods: the LOD support interval (Dupuis and Siegmund 1999) and BCI (Broman and Sen 2009). Another approach to quantify uncertainty on QTL location is to sample from the data directly, using methods like bootstrapping and permutations. We evaluated three sampling-based QTL location intervals: parametric bootstraps, parametric permutations, and Bayesian bootstraps (Rubin 1981). See Appendix C for more details on the sampling-based methods. We also note that in practical terms the likelihood-based approaches are appealing because they require significantly less computation due to not involving sampling.

For all methods, we evaluated how the QTL coverage rate (the rate that the estimated interval included the true QTL across simulations) varied with support or nominal probability. For methods with a nominal probability (*e.g.*, 95% confidence interval), ideally the nominal probability matches the observed coverage rate. We also evaluated the methods based on mapping power when strictly requiring the estimated QTL interval to cover the true location. The support level for a LOD support interval with 80% coverage varies with population and statistical threshold, ranging from 1.25 in the CC and balanced CC-RIX to less than 1 (≈0.7) in the unbalanced CC-RIX samples (Figure S6A). The likelihood-based BCI performs similarly to LOD support interval, though power is slightly reduced for all populations (Figure S6B). BCI is also conservative in terms coverage rate because the observed coverage is generally higher than the nominal probability.

Overall, the sampling-based intervals did not perform as well as the likelihood-based intervals for mapping power (Figure S7). The observed coverage rate for sampling-based methods did generally track more closely with the nominal probability (and even matched in some populations) than for BCI. In terms of coverage rate and the median interval width (narrower is better), the likelihood-based intervals generally performed better (Figure 7 excludes Bayesian bootstrap due to poor performance and Figure S8 includes it). For likelihood-based methods, the BCIs were slightly narrower than the LOD support intervals, but the LOD support intervals had slightly better coverage. For sampling-based intervals, parametric bootstraps and permutations outperformed Bayesian bootstraps.

We note that fitting the founder haplotype model can result in unstable associations when there are very rare alleles, which can occur when a founder allele has been lost at a locus (Keele *et al.*

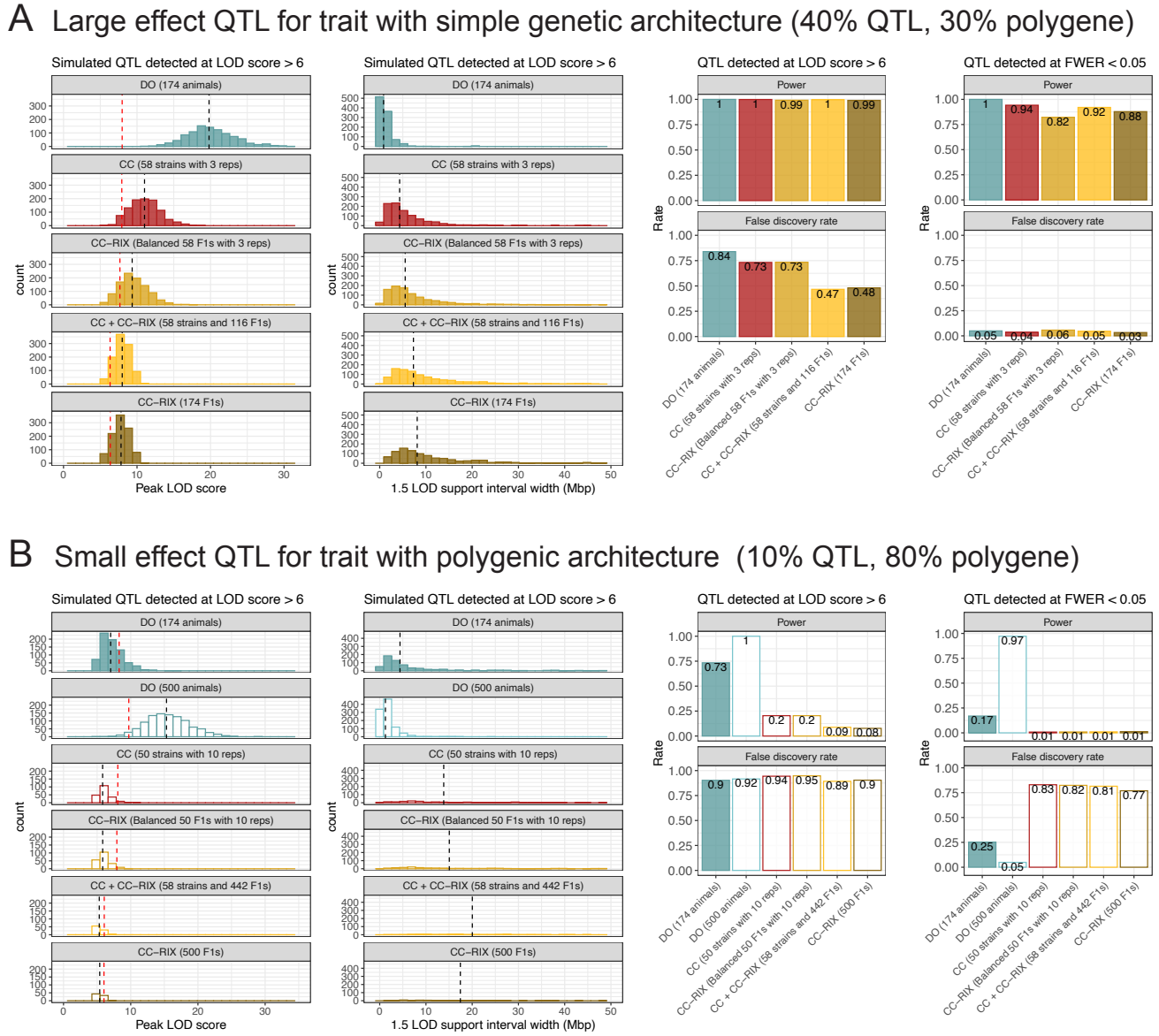

Figure 6 Performance of QTL mapping in data simulated for CC, CC-RIX, and DO populations from two scenarios: (A) large effect QTL in a trait with relatively simple genetic architecture (40% QTL and 30% polygene) in 174 mice and (B) small effect QTL in a trait with highly polygenic genetic architecture (10% QTL and 80% polygene) in 500 mice (as well as 174 DO mice for comparison). For each scenario, 1,000 QTLs were simulated and mapped using a lenient significance threshold of LOD score > 6 and summarized as histograms of maximum LOD score and 1.5 LOD support intervals. Black vertical dashed lines represent medians and the red vertical dashed lines represent the 95% significance thresholds. Power and false discovery rates based on lenient and 95% significance thresholds.

2018; Hsiao *et al.* 2020). We speculate that the likelihood-based intervals could be susceptible to these problematic loci, producing overly certain, *i.e.*, narrow, QTL intervals, whereas a sampling approach could more clearly highlight a QTL as problematic.

Homozygosity increases power to map additive QTL in the CC relative to the CC-RIX and DO

For the simulations reported above for mapping power, we strictly scaled the effects of an additive QTL to a specified proportion of the total variation with respect to the observed mapping population. Consider the case of a causal allele that is rare in one sample population but not another; this scaling implicitly increases the effect of the QTL in the population where the allele is rare to equalize the variation explained across populations. More broadly speaking, scaling with respect to variation in the sample population will increase the QTL effect for deviations from balanced homozygous allele frequencies at the QTL, which will occur when there are heterozygous animals rather than homozygous, alleles that are rare, and even imbalanced allelic series, *i.e.*, how the QTL's functional alleles are distributed among the founder strains (Crouse *et al.* 2020). Allele frequencies at loci will vary across the sample populations of CC, DO, and CC-RIX, but the most striking difference is the inbred homozygosity of the CC compared to the large-scale heterozygosity of the DO and CC-RIX.

We evaluated how these features of a population would affect mapping power by again performing 1,000 simulations from the

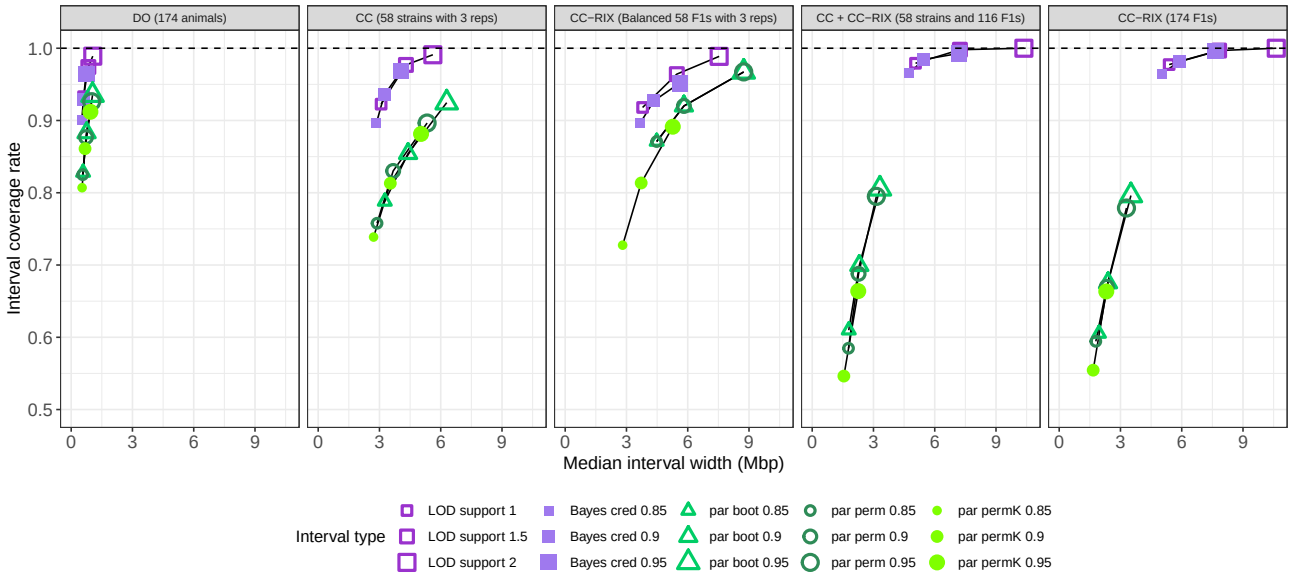

Figure 7 Performance of QTL location interval estimation across data simulated for CC, CC-RIX, and DO populations in terms of coverage rate (y-axis) and interval width (x-axis). Intervals are summarized over 1,000 QTLs simulated in 174 animals for each sample population. For parametric permutation intervals with kinship, 200 samples were generated; for the other sampling-based interval types, 1,000 samples were used. Purple shapes represent likelihood-based procedures and green shapes represent sampling-based procedures. Solid lines connect summaries from the same procedure but with differing support levels, with increasing support indicated by larger symbols. Horizontal dashed line representing complete coverage included for reference. Bayesian bootstrap summaries were excluded due to poor performance, but can be seen in Figure S8.

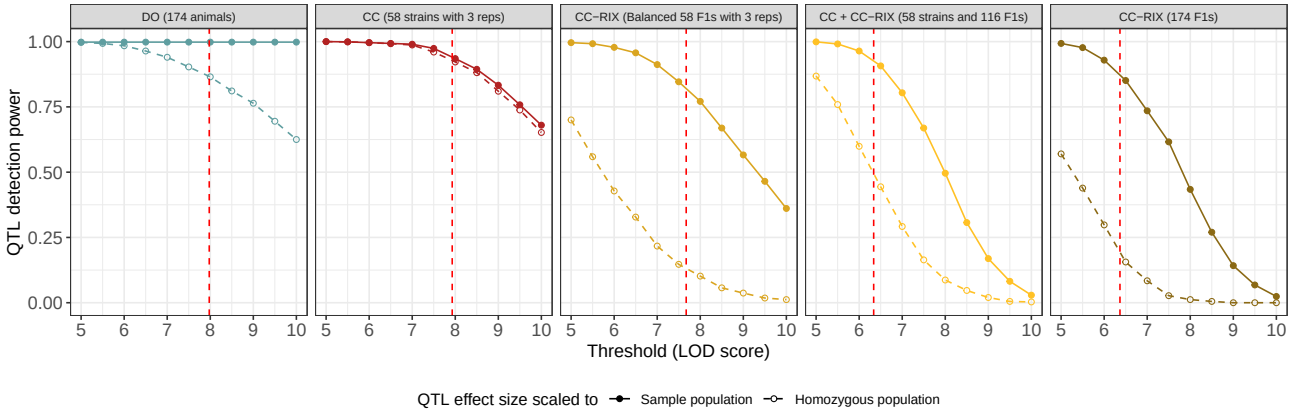

Figure 8 Comparison of QTL mapping power between data simulated for CC, CC-RIX, and DO populations with QTL effects scaled with respect to the sample population being mapped (solid dots and lines) or a homozygous population with balanced representation of alleles (open dots and dashed lines). Red vertical dashed lines represent the 95% significance thresholds for each population.

large QTL scenario (174 mice with 40% QTL and 30% polygene), but now scaling the QTL effect size with respect to a population that at the QTL is homozygous and has balanced allele frequencies (Figure 8). Notably, mapping power in the CC is only slightly reduced, representing a reduction in QTL effect size due to slight imbalances in allele frequency. The DO experiences a noticeable reduction in mapping power, now slightly lower than the CC, due to large-scale heterozygosity. Power in the CC-RIX populations is penalized the most, likely due to the combined effects of heterozygosity and greater imbalance in allele frequencies and population structure.

These findings suggest there is some cause for optimism when mapping in the CC, where an additive QTL effect is likely to be maximized within the sample population. It is important to note that deviations from additivity would lessen this effect. We emphasize that despite the CC's potential to have greater power in some contexts, the benefits of this feature is limited by the reduced number of unique genomes compared to the DO. Large sample populations of DO are by far the best option for mapping highly polygenic traits.

Using the musppr R package for future studies

We designed musppr to be reusable, allowing researchers to input genetic data from sample populations of CC, CC-RIX, and DO they have generated, and thus tailoring findings to specific studies. Its functions are amenable to being run in parallel on a computing

cluster, allowing deeper evaluations of experimental performance, such as mapping power across more effect size settings, which could be useful for proposals and when planning experiments. The broad findings reported here, such as the value of genetic replicates for estimating heritability, are also largely valid when extrapolating to non-mouse MPPs and all recombinant populations. We do caveat that to use musppr to analyze genetic data from non-mouse MPPs, functions may need to be expanded. For example, if the model underlying the QTL mapping analysis requires population-specific features that are not available in the model fit by the qtl2 R package, the mapping musppr function would need to be adjusted.

Conclusions

Here we evaluated the performance of three related genetically diverse mouse MPPs, the CC, CC-RIX, and DO, in estimating heritability and mapping QTLs, commonly used genetic analyses for the study of complex traits. Our findings provide examples of best practices for researchers designing studies with these population resources, such as using the ASV form of the kinship matrix for heritability estimation. More broadly, this work reveals the relative strengths of these populations. Replicate mice in the CC and CC-RIX samples result in more efficient estimation of heritability, potentially offering more precise estimates from far fewer animals than would be required in the DO. The CC and CC-RIX can be powerful tools for genetic mapping when the QTL effect is large ($\geq$ 40%) and the genetic architecture is fairly simple, but as the trait becomes more polygenic, only large sample populations of DO are likely to be well-powered for QTL mapping. The complex population structure of the CC-RIX reduces mapping power, but does enable more accurate estimation of additive heritability. Furthermore, the CC-RIX can be used to detect parent-of-origin effects using reciprocal F1 designs (Oreper *et al.* 2018; Sun *et al.* 2021), though we do not investigate these approaches here. These key principles should also extend to MPPs (and more broadly any mapping population) of other organisms with similar experimental features (*e.g.*, genetic replicates, inbred/outbred).

These results emphasize the complementary nature of these populations for joint analyses. For example, even though CC and CC-RIX sample populations are unlikely to be sufficiently powered to dissect highly polygenic traits, a leniently detected result could confirm a QTL stringently detected in a larger DO sample population; furthermore, replicable CC and CC-RIX mice that possess key alleles of the QTL identified in the DO could be used for follow-up mechanistic studies. Taken together, they represent a flexible and powerful MPP resource for a new generation of complex trait studies.

Data availability

All analyses were performed using the R statistical programming language (R Core Team 2022). We wrote the musppr R package to perform all simulation and subsequent analysis, which is available at https://github.com/gkeele/musppr. Founder haplotype data for all populations used in this study, a fixed version of musppr, and R code used to generate the simulated data, figures, and reported results can be found at https://doi.org/10.6084/m9.figshare.20560821.

Acknowledgments

We acknowledge Martin T. Ferris and William Valdar, both of the University of North Carolina at Chapel Hill, and Gary A. Churchill of the Jackson Laboratory for discussions related to this work. We thank Callan O'Connor of the Jackson Laboratory and Paul L. Maurizio of the University of Chicago for providing feedback on an early draft manuscript.

Funding

This work was supported by grants from the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH): F32GM134599 and R01GM070683.

Conflicts of interest

The author declares no conflicts of interest.

Literature cited

Al-Barghouthi BM, Mesner LD, Calabrese GM, Brooks D, Tommasini SM, Bouxsein ML, Horowitz MC, Rosen CJ, Nguyen K, Haddox S *et al*. 2021. Systems genetics in diversity outbred mice inform BMD GWAS and identify determinants of bone strength. Nature Communications. 12:3408.

Aydin S, Pham DT, Zhang T, Keele GR, Skelly DA, Pankratz M, Choi T, Gygi SP, Reinholdt LG, Baker CL *et al*. 2022. Genetic dissection of the pluripotent proteome through multi-omics data integration. bioRxiv. .

Aylor DL, Valdar W, Foulds-Mathes W, Buus RJ, Verdugo RA, Baric RS, Ferris MT, Frelinger JA, Heise M, Frieman MB *et al*. 2011. Genetic analysis of complex traits in the emerging Collaborative Cross. Genome Research. 21:1213–1222.

Bandillo N, Raghavan C, Muyco PA, Sevilla MAL, Lobina IT, Dilla-Ermita CJ, Tung CW, McCouch S, Thomson M, Mauleon R *et al*. 2013. Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice. 6:11.

Broman KW, Gatti DM, Simecek P, Furlotte NA, Prins P, Sen Yandell BS, Churchill GA. 2019. R/qtl2: Software for Mapping Quantitative Trait Loci with High-Dimensional Data and Multiparent Populations. Genetics. 211:495–502.

Broman KW, Sen S. 2009. *A Guide to QTL Mapping with R/qtl*. Statistics for Biology and Health. Springer New York. New York, NY.

Cheng R, Parker CC, Abney M, Palmer AA. 2013. Practical Considerations Regarding the Use of Genotype and Pedigree Data to Model Relatedness in the Context of Genome-Wide Association Studies. G3 Genes | Genomes | Genetics. 3:1861–1867.

Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, Raghupathy N, Svenson KL, Churchill GA, Gygi SP. 2016. Defining the consequences of genetic variation on a proteome-wide scale. Nature. 534:500–505.

Churchill GA, Doerge RW. 1994. Empirical threshold values for quantitative trait mapping. Genetics. 138:963–971.

Churchill GA, Gatti DM, Munger SC, Svenson KL. 2012. The diversity outbred mouse population. Mammalian Genome. 23:713–718.

Collaborative Cross Consortium. 2012. The Genome Architecture of the Collaborative Cross Mouse Genetic Reference Population. Genetics. 190:389–401.

Covarrubias-Pazaran G. 2016. Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. PLOS ONE. 11:e0156744.

Crouse WL, Kelada SNP, Valdar W. 2020. Inferring the Allelic Series at QTL in Multiparental Populations. Genetics. 216:957–983.

Cubillos FA, Parts L, Salinas F, Bergström A, Scovacricchi E, Zia A, Illingworth CJR, Mustonen V, Ibstedt S, Warringer J *et al*. 2013. High-Resolution Mapping of Complex Traits with a Four-Parent Advanced Intercross Yeast Population. Genetics. 195:1141–1155.

de Koning DJ, McIntyre LM. 2017. Back to the Future: Multiparent Populations Provide the Key to Unlocking the Genetic Basis of Complex Traits. G3 Genes | Genomes | Genetics. 7:1617–1618.

Dell'Acqua M, Gatti DM, Pea G, Cattonaro F, Coppens F, Magris G, Hlaing AL, Aung HH, Nelissen H, Baute J *et al*. 2015. Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea mays. Genome Biology. 16:167.

Dudbridge F, Koeleman BP. 2004. Efficient Computation of Significance Levels for Multiple Associations in Large Studies of Correlated Data, Including Genomewide Association Studies. The American Journal of Human Genetics. 75:424–435.

Dupont MSJ, Guillemot V, Campagne P, Serafini N, Marie S, Montagutelli X, Di Santo JP, Vosshenrich CAJ. 2021. Host genetic control of natural killer cell diversity revealed in the Collaborative Cross. Proceedings of the National Academy of Sciences. 118:e2018834118.

Dupuis J, Siegmund D. 1999. Statistical Methods for Mapping Quantitative Trait Loci From a Dense Set of Markers. Genetics. 151:373–386.

Eldridge R, Osorio D, Amstalden K, Edwards C, Young CR, Cai JJ, Konganti K, Hillhouse A, Threadgill DW, Welsh CJ *et al*. 2020. Antecedent presentation of neurological phenotypes in the Collaborative Cross reveals four classes with complex sex-dependencies. Scientific Reports. 10:7918.

Endelman JB, Jannink JL. 2012. Shrinkage Estimation of the Realized Relationship Matrix. G3 Genes | Genomes | Genetics. 2:1405–1413.

Falconer D, Mackay T. 1996. *Introduction to Quantitative Genetics*. Longmans Green. Harlow, Essex, UK. fourth edition.

Feldmann MJ, Piepho HP, Knapp SJ. 2022. Average semivariance directly yields accurate estimates of the genomic variance in complex trait analyses. G3 Genes | Genomes | Genetics. 12:jkac080.

Ferris MT, Aylor DL, Bottomly D, Whitmore AC, Aicher LD, Bell TA, Bradel-Tretheway B, Bryan JT, Buus RJ, Gralinski LE *et al*. 2013. Modeling Host Genetic Regulation of Influenza Pathogenesis in the Collaborative Cross. PLoS Pathogens. 9:e1003196.

French JE, Gatti DM, Morgan DL, Kissling GE, Shockley KR, Knudsen GA, Shepard KG, Price HC, King D, Witt KL *et al*. 2015. Diversity Outbred Mice Identify Population-Based Exposure Thresholds and Genetic Factors that Influence Benzene-Induced Genotoxicity. Environmental Health Perspectives. 123:237–245.

Gatti DM, Svenson KL, Shabalin A, Wu LY, Valdar W, Simecek P, Goodwin N, Cheng R, Pomp D, Palmer A *et al*. 2014. Quantitative Trait Locus Mapping Methods for Diversity Outbred Mice. G3 Genes | Genomes | Genetics. 4:1623–1633.

Gerdes Gyuricza I, Chick JM, Keele GR, Deighan AG, Munger SC, Korstanje R, Gygi SP, Churchill GA. 2022. Genome-wide transcript and protein analysis highlights the role of protein homeostasis in the aging mouse heart. Genome Research. p. genome;gr.275672.121v2.

Gould RL, Craig SW, McClatchy S, Churchill GA, Pazdro R. 2021. Quantitative trait mapping in Diversity Outbred mice identifies novel genomic regions associated with the hepatic glutathione redox system. Redox Biology. 46:102093.

Gralinski LE, Ferris MT, Aylor DL, Whitmore AC, Green R, Frieman MB, Deming D, Menachery VD, Miller DR, Buus RJ *et al*. 2015. Genome Wide Identification of SARS-CoV Susceptibility Loci Using the Collaborative Cross. PLOS Genetics. 11:e1005504.

Haley CS, Knott SA. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity. 69:315–324.

Hampton BK, Jensen KL, Whitmore AC, Linnertz CL, Maurizio P, Miller DR, Morrison CR, Noll KE, Plante KS, Shaw GD *et al*. 2021. Genetic regulation of homeostatic immune architecture in the lungs of collaborative cross mice. bioRxiv. .

Hsiao K, Noble C, Pitman W, Yadav N, Kumar S, Keele GR, Terceros A, Kanke M, Conniff T, Cheleuitte-Nieves C *et al*. 2020. A Thalamic Orphan Receptor Drives Variability in Short-Term Memory. Cell. 183:522–536.e19.

Keele GR, Crouse WL, Kelada SNP, Valdar W. 2019. Determinants of QTL Mapping Power in the Realized Collaborative Cross. G3 Genes | Genomes | Genetics. 9:1707–1727.

Keele GR, Prokop JW, He H, Holl K, Littrell J, Deal A, Francic S, Cui L, Gatti DM, Broman KW *et al*. 2018. Genetic Fine-Mapping and Identification of Candidate Genes and Variants for Adiposity Traits in Outbred Rats: Mapping Adiposity Traits in Outbred Rats. Obesity. 26:213–222.

Keele GR, Quach BC, Israel JW, Chappell GA, Lewis L, Safi A, Simon JM, Cotney P, Crawford GE, Valdar W *et al*. 2020. Integrative QTL analysis of gene expression and chromatin accessibility identifies multi-tissue patterns of genetic regulation. PLOS Genetics. 16:e1008537.

Keele GR, Zhang T, Pham DT, Vincent M, Bell TA, Hock P, Shaw GD, Paulo JA, Munger SC, Pardo-Manuel de Villena F *et al*. 2021. Regulation of protein abundance in genetically diverse mouse populations. Cell Genomics. 1:100003.

Kelada SNP, Aylor DL, Peck BCE, Ryan JF, Tavarez U, Buus RJ, Miller DR, Chesler EJ, Threadgill DW, Churchill GA *et al*. 2012. Genetic Analysis of Hematological Parameters in Incipient Lines of the Collaborative Cross. G3 Genes | Genomes | Genetics. 2:157–165.

Kelada SNP, Carpenter DE, Aylor DL, Chines P, Rutledge H, Chesler EJ, Churchill GA, Pardo-Manuel de Villena F, Schwartz DA, Collins FS. 2014. Integrative Genetic Analysis of Allergic Inflammation in the Murine Lung. American Journal of Respiratory Cell and Molecular Biology. 51:436–445.

Keller MP, Gatti DM, Schueler KL, Rabaglia ME, Stapleton DS, Simecek P, Vincent M, Allen S, Broman AT, Bacher R *et al*. 2018. Genetic Drivers of Pancreatic Islet Function. Genetics. 209:335–356.

Keller MP, Rabaglia ME, Schueler KL, Stapleton DS, Gatti DM, Vincent M, Mitok KA, Wang Z, Ishimura T, Simonett SP *et al*. 2019. Gene loci associated with insulin secretion in islets from nondiabetic mice. Journal of Clinical Investigation. 129:4419–4432.

Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R. 2009. A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in Arabidopsis thaliana. PLoS Genetics. 5:e1000551.

Lander ES, Botstein D. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. 121:185–199.

Linke V, Overmyer KA, Miller IJ, Brademan DR, Hutchins PD, Trujillo EA, Reddy TR, Russell JD, Cushing EM, Schueler KL *et al*. 2020. A large-scale genome–lipid association map guides lipid identification. Nature Metabolism. 2:1149–1162.

Long AD, Macdonald SJ, King EG. 2014. Dissecting complex traits using the Drosophila Synthetic Population Resource. Trends in Genetics. 30:488–495.

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

Manichaikul A, Dupuis J, Sen Broman KW. 2006. Poor Performance of Bootstrap Confidence Intervals for the Location of a

Quantitative Trait Locus. Genetics. 174:481–489.

Martínez O, Curnow RN. 1992. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theoretical and Applied Genetics. 85:480–488.

Morgan AP, Welsh CE. 2015. Informatics resources for the Collaborative Cross and related mouse populations. Mammalian Genome. 26:521–539.

Mosedale M, Cai Y, Eaddy JS, Corty RW, Nautiyal M, Watkins PB, Valdar W. 2019. Identification of Candidate Risk Factor Genes for Human Idelalisib Toxicity Using a Collaborative Cross Approach. Toxicological Sciences. 172:265–278.

Mosedale M, Cai Y, Eaddy JS, Kirby PJ, Wolenski FS, Dragan Y, Valdar W. 2021. Human-relevant mechanisms and risk factors for TAK-875-Induced liver injury identified via a gene pathway-based approach in Collaborative Cross mice. Toxicology. 461:152902.

Mosedale M, Kim Y, Brock WJ, Roth SE, Wiltshire T, Scott Eaddy J, Keele GR, Corty RW, Xie Y, Valdar W *et al*. 2017. Candidate Risk Factors and Mechanisms for Tolvaptan-Induced Liver Injury Are Identified Using a Collaborative Cross Approach. Toxicological Sciences. p. kfw269.

Noble LM, Chelo I, Guzella T, Afonso B, Riccardi DD, Ammerman P, Dayarian A, Carvalho S, Crist A, Pino-Querido A *et al*. 2017. Polygenicity and Epistasis Underlie Fitness-Proximal Traits in the *Caenorhabditis elegans* Multiparental Experimental Evolution (CeMEE) Panel. Genetics. 207:1663–1685.

Oreper D, Schoenrock SA, McMullan R, Ervin R, Farrington J, Miller DR, de Villena FPM, Valdar W, Tarantino LM. 2018. Reciprocal F1 Hybrids of Two Inbred Mouse Strains Reveal Parent-of-Origin and Perinatal Diet Effects on Behavior and Expression. G3 Genes | Genomes | Genetics. 8:3447–3468.

Orgel K, Smeekens JM, Ye P, Fotsch L, Guo R, Miller DR, Pardo-Manuel de Villena F, Burks AW, Ferris MT, Kulis MD. 2019. Genetic diversity between mouse strains allows identification of the CC027/GeniUnc strain as an orally reactive model of peanut allergy. Journal of Allergy and Clinical Immunology. 143:1027–1037.e7.

Ouellette AR, Neuner SM, Dumitrescu L, Anderson LC, Gatti DM, Mahoney ER, Bubier JA, Churchill G, Peters L, Huentelman MJ *et al*. 2020. Cross-Species Analyses Identify Dlgap2 as a Regulator of Age-Related Cognitive Decline and Alzheimer's Dementia. Cell Reports. 32:108091.

Patterson HD, Thompson R. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika. 58:545–554.

Philip VM, Sokoloff G, Ackert-Bicknell CL, Striz M, Branstetter L, Beckmann MA, Spence JS, Jackson BL, Galloway LD, Barker P *et al*. 2011. Genetic analysis in the Collaborative Cross breeding population. Genome Research. 21:1223–1238.

R Core Team. 2022. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rogala AR, Morgan AP, Christensen AM, Gooch TJ, Bell TA, Miller DR, Godfrey VL, de Villena FPM. 2014. The Collaborative Cross as a Resource for Modeling Human Disease: CC011/Unc, a New Mouse Model for Spontaneous Colitis. Mammalian Genome. 25:95–108.

Rubin DB. 1981. The bayesian bootstrap. The Annals of Statistics. 9:130–134.

Schoenrock SA, Oreper D, Farrington J, McMullan RC, Ervin R, Miller DR, Pardo-Manuel de Villena F, Valdar W, Tarantino LM. 2018. Perinatal nutrition interacts with genetic background to alter behavior in a parent-of-origin-dependent manner in adult Collaborative Cross mice. Genes, Brain and Behavior. 17:e12438.

Schäfer A, Leist SR, Gralinski LE, Martinez DR, Winkler ES, Okuda K, Hawkins PE, Gully KL, Graham RL, Scobey DT *et al*. 2022. A Multitrait Locus Regulates Sarbecovirus Pathogenesis. preprint. Microbiology.

Scoggin K, Lynch R, Gupta J, Nagarajan A, Sheffield M, Elsaadi A, Bowden C, Aminian M, Peterson A, Adams LG *et al*. 2022. Genetic background influences survival of infections with Salmonella enterica serovar Typhimurium in the Collaborative Cross. PLOS Genetics. 18:e1010075.

Shorter JR, Odet F, Aylor DL, Pan W, Kao CY, Fu CP, Morgan AP, Greenstein S, Bell TA, Stevans AM *et al*. 2017. Male Infertility Is Responsible for Nearly Half of the Extinction Observed in the Mouse Collaborative Cross. Genetics. 206:557–572.

Sigmon JS, Blanchard MW, Baric RS, Bell TA, Brennan J, Brockmann GA, Burks AW, Calabrese JM, Caron KM, Cheney RE *et al*. 2020. Content and Performance of the MiniMUGA Genotyping Array: A New Tool To Improve Rigor and Reproducibility in Mouse Research. Genetics. 216:905–930.

Skelly DA, Czechanski A, Byers C, Aydin S, Spruce C, Olivier C, Choi K, Gatti DM, Raghupathy N, Keele GR *et al*. 2020. Mapping the Effects of Genetic Variation on Chromatin State and Gene Expression Reveals Loci That Control Ground State Pluripotency. Cell Stem Cell. 27:459–469.e8.

Smith CM, Proulx MK, Lai R, Kiritsy MC, Bell TA, Hock P, Pardo-Manuel de Villena F, Ferris MT, Baker RE, Behar SM *et al*. 2019. Functionally Overlapping Variants Control Tuberculosis Susceptibility in Collaborative Cross Mice. mBio. 10:e02791–19.

Srivastava A, Morgan AP, Najarian ML, Sarsani VK, Sigmon JS, Shorter JR, Kashfeen A, McMullan RC, Williams LH, Giusti-Rodríguez P *et al*. 2017. Genomes of the Mouse Collaborative Cross. Genetics. 206:537–556.

Sun KY, Oreper D, Schoenrock SA, McMullan R, Giusti-Rodríguez P, Zhabotynsky V, Miller DR, Tarantino LM, Pardo-Manuel de Villena F, Valdar W. 2021. Bayesian modeling of skewed X inactivation in genetically diverse mice identifies a novel *Xce* allele associated with copy number changes. Genetics. 218:iyab034.

Svenson KL, Gatti DM, Valdar W, Welsh CE, Cheng R, Chesler EJ, Palmer AA, McMillan L, Churchill GA. 2012. High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population. Genetics. 190:437–447.

Takemon Y, Chick JM, Gerdes Gyuricza I, Skelly DA, Devuyst O, Gygi SP, Churchill GA, Korstanje R. 2021. Proteomic and transcriptomic profiling reveal different aspects of aging in the kidney. eLife. 10:e62585.

The Complex Trait Consortium. 2004. The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nature Genetics. 36:1133–1137.

Threadgill DW, Miller DR, Churchill GA, de Villena FPM. 2011. The Collaborative Cross: A Recombinant Inbred Mouse Population for the Systems Genetic Era. ILAR Journal. 52:24–31.

Tovar A, Smith GJ, Nalesnik MB, Thomas JM, McFadden KM, Harkema JR, Kelada SNP. 2022. A Locus on Chromosome 15 Contributes to Acute Ozone-Induced Lung Injury in Collaborative Cross Mice. American Journal of Respiratory Cell and Molecular Biology. pp. rcmb.2021–0326OC.

Valdar W, Flint J, Mott R. 2006. Simulating the Collaborative Cross: Power of Quantitative Trait Loci Detection and Mapping Resolution in Large Sets of Recombinant Inbred Strains of Mice. Genetics. 172:1783–1797.

Visscher PM, Thompson R, Haley CS. 1996. Confidence Intervals in QTL Mapping by Bootstrapping. Genetics. 143:1013–1020.

Woods LCS, Mott R. 2017. Heterogeneous Stock Populations for

Analysis of Complex Traits, In: Schughart K, Williams RW, editors, *Systems Genetics*, Springer New York. New York, NY. volume 1488. pp. 31–44. Series Title: Methods in Molecular Biology.

Yam P, Albright J, VerHague M, Gertz ER, Pardo-Manuel de Villena F, Bennett BJ. 2021. Genetic Background Shapes Phenotypic Response to Diet for Adiposity in the Collaborative Cross. Frontiers in Genetics. 11:615012.

Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F. 2007. On the subspecific origin of the laboratory mouse. Nature Genetics. 39:1100–1107.

Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, Bonhomme F, Yu AHT, Nachman MW, Pialek J *et al*. 2011. Subspecific origin and haplotype diversity in the laboratory mouse. Nature Genetics. 43:648–655.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. 2014. Advantages and pitfalls in the application of mixed-model association methods. Nature Genetics. 46:100–106.

Zhang J, Teh M, Kim J, Eva MM, Cayrol R, Meade R, Nijnik A, Montagutelli X, Malo D, Jaubert J. 2019. A loss-of-function mutation in the integrin alpha l (itgal) gene contributes to susceptibility to salmonella enterica serovar typhimurium infection in collaborative cross strain cc042. Infection and immunity. 88:e00656–19.

Zhang T, Keele GR, Gyuricza IG, Vincent M, Brunton C, Bell TA, Hock P, Shaw GD, Munger SC, de Villena FPM *et al*. 2022. Multiomics analysis identifies drivers of protein phosphorylation. bioRxiv. .

Appendix A

Processing founder haplotype data

Our initial populations were 116 CC mice (2 animals per 58 strains) genotyped on MiniMUGA (11,000 markers), 192 DO mice genotyped on MegaMUGA (57,000 markers), and 500 DO mice genotyped on GigaMUGA (143,000 markers). We used the qtl2 R package to impute all populations to the same 64,000 loci. Though the shared set of loci far exceeds the mapping resolution of the CC, it is appealing to compare populations based on the same set of loci.

The input data for the DO are scaled haplotype dosages, reflecting an additive model rather than the full 36-state space of outbred diplotypes. Denser genotype arrays result in greater certainty in founder haplotype pair (*i.e.*, diplotype) estimation. To keep this from influencing our findings, we imputed all loci based on most likely haplotype.

CC individuals and strains: Consider diplotype probability vector for CC mouse i at the locus p:

$$\tilde{\mathbf{a}}_{i,p}^{\mathrm{T}} = [0.1\ 0.05\ 0.05\ 0.1\ 0.6\ 0.02\ 0.03\ 0.05]$$

where elements represent the homozygous diplotype probabilities for the founder haplotypes ordered as AJ, B6, 129, NOD, NZO, CAST, PWK, and WSB. For this example, we impute the mouse as having the NZO diplotype:

$$\mathbf{a}_{i,p}^{\mathrm{T}} = [0\ 0\ 0\ 0\ 1\ 0\ 0\ 0]$$

For the CC, we performed this step both at the individual-level and strain-level. For strain-level, we first averaged the diplotype probabilities at all loci between the two individuals from the same strain before doing imputation of most likely diplotype, thus smoothing out loci with segregating variation within a CC strain.

CC-RIX: We derived founder haplotype data for the $\binom{58}{2} = 1,653$ possible F1s given 58 CC strains (ignoring reciprocal F1s). Consider CC strains j and k that have the B6 and PWK alleles at locus p, respectively:

$$\mathbf{a}_{j,p}^{\mathrm{T}} = [0\ 1\ 0\ 0\ 0\ 0\ 0\ 0]$$

$$\mathbf{a}_{k,p}^{\mathrm{T}} = [0\ 0\ 0\ 0\ 0\ 0\ 1\ 0]$$

The F1's scaled haplotype count at the locus p is calculated by the taking element-wise averages:

$$\mathbf{a}_{jk,p}^{\mathrm{T}} = [0\ 0.5\ 0\ 0\ 0\ 0\ 0.5\ 0]$$

This process was repeated across all loci for all pairwise combinations of CC strains.

DO: Selecting most likely haplotypes for the DO samples is more complicated than the CC due to input being scaled haplotype dosages rather than 36-state probibilities. Consider the scaled haplotype dosage vector for DO mouse i at the locus p, we calculate $\delta_{i,p}^{(1)} = \min(\tilde{\mathbf{a}}_{i,p} - 1)$ and $\delta_{i,p}^{(0.5)} = \min(\tilde{\mathbf{a}}_{i,p} - 0.5)$. If $\delta_{i,p}^{(1)} < \delta_{i,p}^{(0.5)}$, the dosages suggest mouse i is homozygous at locus p and we set the allele with maximum probability from $\tilde{\mathbf{a}}_{i,p}$ to 1 and all others to 0. If $\delta_{i,p}^{(1)} > \delta_{i,p}^{(0.5)}$, the dosages suggest mouse i is heterozygous at locus p and for $\mathbf{a}_{i,p}$ we set the alleles with highest and second highest probability from $\tilde{\mathbf{a}}_{i,p}$ to 0.5 and all others to 0.

Appendix B

Simulating data for heritability estimation

We extended our simulation approach for QTL mapping (Keele *et al.* 2019) to heritability, built from the model specified in Eq 1. We generate simulation t using the following model:

$$\mathbf{y}_{(t)}^{\mathrm{sim}} = s_{(t)}^{\mathbf{u}} * \mathbf{u}_{(t)}^{*} + s_{(t)}^{\boldsymbol{\varepsilon}} * \boldsymbol{\varepsilon}_{(t)}^{*} \tag{7}$$

An intercept can be included but by default we set it to 0. We randomly sample draws of $\mathbf{u}_{(t)}^{*}$ and $\boldsymbol{\varepsilon}_{(t)}^{*}$ according to $\mathbf{u}_{(t)}^{*} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ and $\boldsymbol{\varepsilon}_{(t)}^{*} \sim \mathcal{N}(0, 1)$. To control the relative contributions of a polygene effect ($\mathbf{u}$) at a specified proportion of total variation ϕ^2, we scale the random draws accordingly:

$$s_{(t)}^{\mathbf{u}} = \left(\frac{n\phi^2}{(n-1)V(\mathbf{u}_{(t)}^{*})} \right)^{\frac{1}{2}}$$

$$s_{(t)}^{\boldsymbol{\varepsilon}} = \left(\frac{n(1-\phi^2)}{(n-1)V(\boldsymbol{\varepsilon}_{(t)}^{*})} \right)^{\frac{1}{2}}$$

where $V(.)$ returns the unbiased sample variance of its argument vector. For simulations with two components of heritability, we used the same approach but with respect to Eq 3, now with two polygenic effects controlled at ϕ^2_{add} and ϕ^2_{rep}.

Appendix C

Estimating confidence intervals for the location of a QTL

We estimated likelihood-based intervals for QTL location using LOD support intervals (Dupuis and Siegmund 1999) and BCIs (Broman and Sen 2009), using the qtl2 R package (Broman *et al.* 2019). These approaches are fast and efficient because they do not involve a sampling process, and are instead summarized directly from the LOD score profile around a detected QTL.

Sampling-based approaches are slow but statistically appealing because they incorporate uncertainty through the sampling process. We evaluated four related sampling-based procedures.

Parametric bootstraps: The sampling process for parametric bootstraps involves simulating data from the QTL model (Eq 6) at the detected locus. We fit the model to generate parametric bootstrap sample t:

$$\mathbf{y}_{(t)}^{\mathrm{parboot}} = \widehat{\mathbf{QTL}}_{[p]} + \boldsymbol{\epsilon}_{(t)} \tag{8}$$

where $\widehat{\mathbf{QTL}}_{[p]}$ is the fit prediction of the QTL (based on eight fixed effect parameters for founder haplotpes) and $\boldsymbol{\epsilon}_{(t)}$ is generated according to $\boldsymbol{\epsilon}_{(t)} \sim \mathcal{N}(0, \widehat{\sigma^2})$ where $\widehat{\sigma^2}$ is the estimated residual variance. We then map all bootstrap samples for the chromosome of the detected QTL and record the peak location. Middle quantile intervals for peak location were then calculated for a specified level, *e.g.*, 95%. Notably, we exclude the polygenic effect to simplify the computational cost of running scans of the bootstrap samples. Incorporating variation from estimating the polygenic effect should make the intervals wider.

Parametric permutations: The sampling process for parametric permutations is similar to parametric bootstraps. To generate parametric permutation sample t, we fit:

$$\mathbf{y}_{(t)}^{\mathrm{parperm}} = \widehat{\mathbf{QTL}}_{[p]} + \pi(\widehat{\boldsymbol{\epsilon}})_{(t)} \tag{9}$$

where $\pi()$ is a function that permutes (randomly reorders) a vector, $\widehat{\epsilon}$ are the fit residuals, and all other terms as previously defined. Subsequent scans of parametric permutations and interval estimation is performed as with the parametric bootstraps.

Parametric permutations with kinship: We also performed the parametric permutation procedure as previously described but with the kinship term fit as part of the subsequent scans, allowing us to assess whether ignoring the kinship effect made a difference.

Bayesian bootstrap: The Bayesian bootstrap procedure is a continuous generalization of nonparametric bootstrapping, *i.e.*, sampling with replacement (Rubin 1981). Instead of sampling with replacement, we generate sample weights for individuals ($\mathbf{w}_{(t)}^{\text{bayesboot}}$), ensuring that all individuals make some contribution to each sample, even if only fractional. The original data are re-scanned for the chromosome with the detected QTL, but now with each sample of weights. Intervals are recorded using the same quantile summary as with the other sampling-based approaches. Initially we included the kinship term but found that when qtl2 was provided a kinship matrix, it disregarded the weights argument. Further work is needed on weighting in the presence of a random term for kinship.

Evaluating the sampling-based procedures represents a significant computational burden because of the additional sampling steps for each simulated QTL. For the 1,000 QTLs that were simulated for the large QTL scenario, we evaluated parametric bootstraps and parametric permutations based on 1,000 samples. For parametric permutations with kinship and Bayesian bootstrap, the number of samples was reduced to 200 because both procedures are computationally more intensive.

Appendix D

Simulating data for QTL mapping

We simulated QTL data using our previous approach in the CC (Keele *et al.* 2019), now extended for CC-RIX and DO data as well. For sample populations of CC or CC-RIX with replicates, we simulated the data as strain/F1 means, reducing the size of the data and making subsequent mapping analysis more computationally efficient. Consider the QTL model specified in Eq 6 with effects on a trait stemming from the QTL, polygene, and random noise. We calculate the expected QTL effect size at the level of strain/F1 means based on the reduction in noise on means. Similar to our simulations of heritability (Appendix B), we generate simulation t for a QTL at p with the following model:

$$\mathbf{y}_{(t)}^{\text{sim}} = s_{(t)}^{\text{QTL}} * \widehat{\mathbf{QTL}}_{[p]} + s_{(t)}^{\mathbf{u}} * \mathbf{u}_{(t)}^{*} + s_{(t)}^{\boldsymbol{\varepsilon}} * \boldsymbol{\varepsilon}_{(t)}^{*} \tag{10}$$

where scaling factors are calculated as

$$s_{(t)}^{\text{QTL}} = \left(\frac{1-(1-\phi_{\text{QTL}}^2-\phi_{\text{u}}^2)/r}{\phi_{\text{QTL}}^2+\phi_{\text{u}}^2}\right)\left(\frac{n\phi_{\text{QTL}}^2}{(n-1)V(\widehat{\mathbf{QTL}}_{[p]})}\right)^{\frac{1}{2}}$$

$$s_{(t)}^{\mathbf{u}} = \left(\frac{1-(1-\phi_{\text{QTL}}^2-\phi_{\text{u}}^2)/r}{\phi_{\text{QTL}}^2+\phi_{\text{u}}^2}\right)\left(\frac{n\phi_{\mathbf{u}}^2}{(n-1)V(\mathbf{u}_{(t)}^{*})}\right)^{\frac{1}{2}}$$

$$s_{(t)}^{\boldsymbol{\varepsilon}} = \left(\frac{n(1-\phi_{\text{QTL}}^2-\phi_{\mathbf{u}}^2)/r}{(n-1)V(\boldsymbol{\varepsilon}_{(t)}^{*})}\right)^{\frac{1}{2}}$$

for the proportions of variation explained by the QTL (ϕ_{QTL}^2) and polygene (ϕ_{u}^2), r is the number of replicates per strain/F1, and all other terms as previously defined.

Supplemental Figures

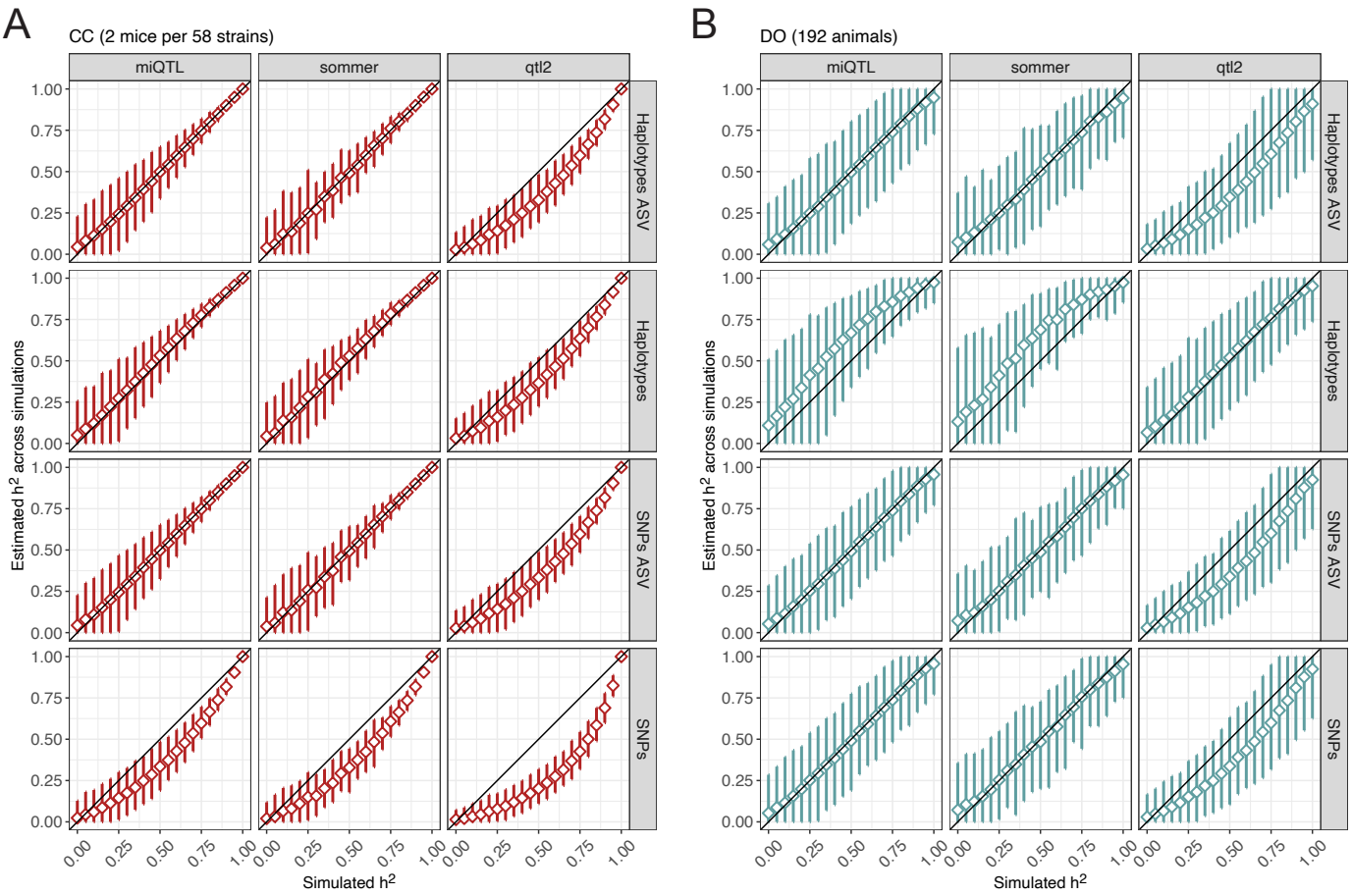

Figure S1 Performance of heritability estimation across kinship matrices (rows) and software packages (columns) in data simulated for (A) 116 CC mice (two per strain) and (B) 192 DO mice. Diamonds represent the mean estimated heritability across simulations from the true heritability. 1,000 simulations were performed with qtl2 and miQTL and 100 were performed for sommer. Vertical line segments represent middle 95% intervals across the simulations. Black lines indicating the heritability estimate is equal to the true value (y = x) included for reference.

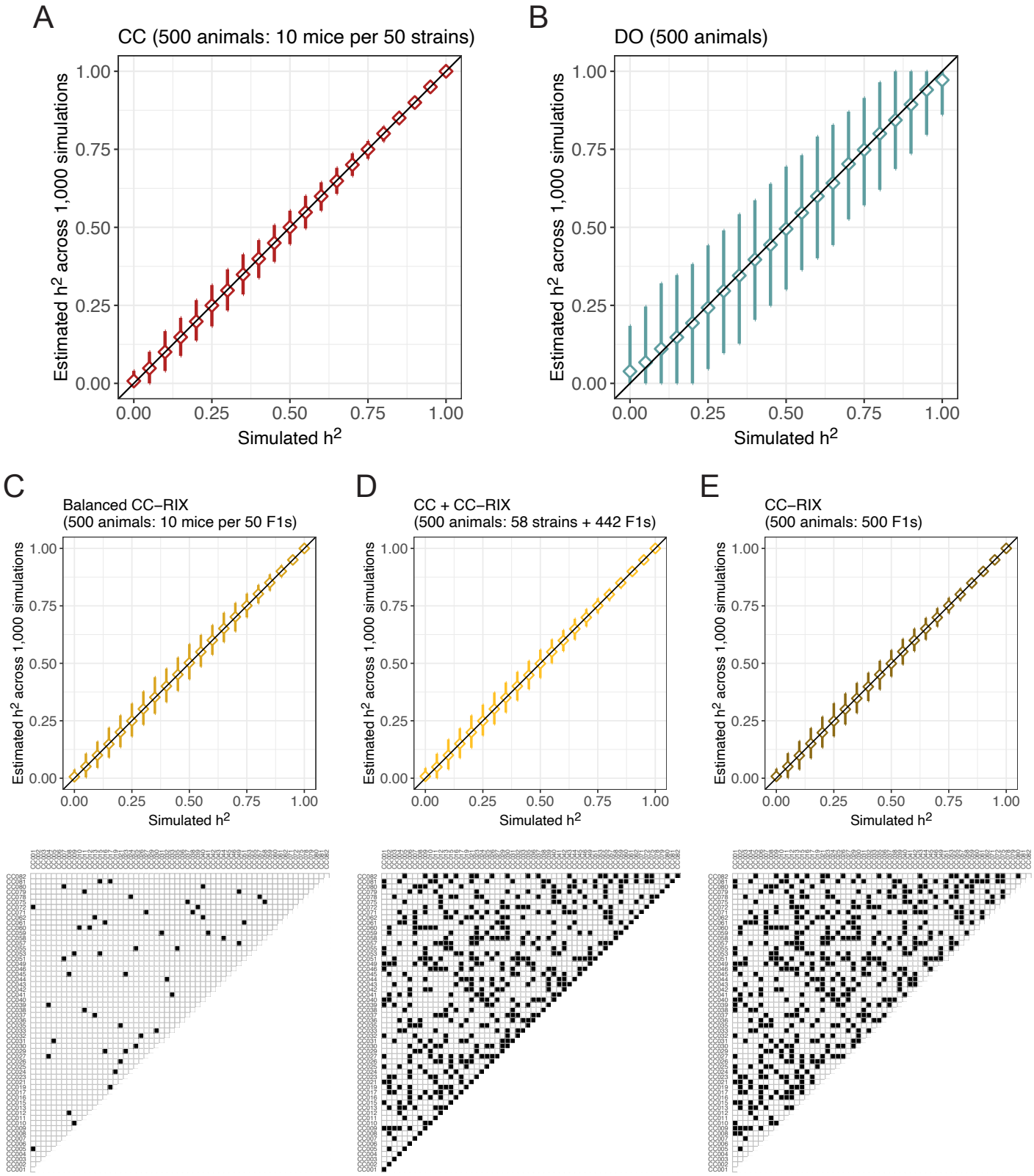

Figure S2 Performance of heritability estimation in data simulated for 500 mice from (A) CC, (B) DO, and (C-E) CC-RIX populations. Diamonds represent the mean estimated heritability across 1,000 simulations from the true heritability. Vertical line segments represent middle 95% intervals across the 1,000 simulations. Black lines indicating the heritability estimate is equal to the true value (y = x) included for reference. Each CC-RIX population (C-E) is represented with a diallel grid of the CC strains and their F1s (ignoring effects of sex chromosomes or parent of origin), with observed strains and F1s as black cells (bottom row). See Figure 3 for results from simulations of 174 mice.

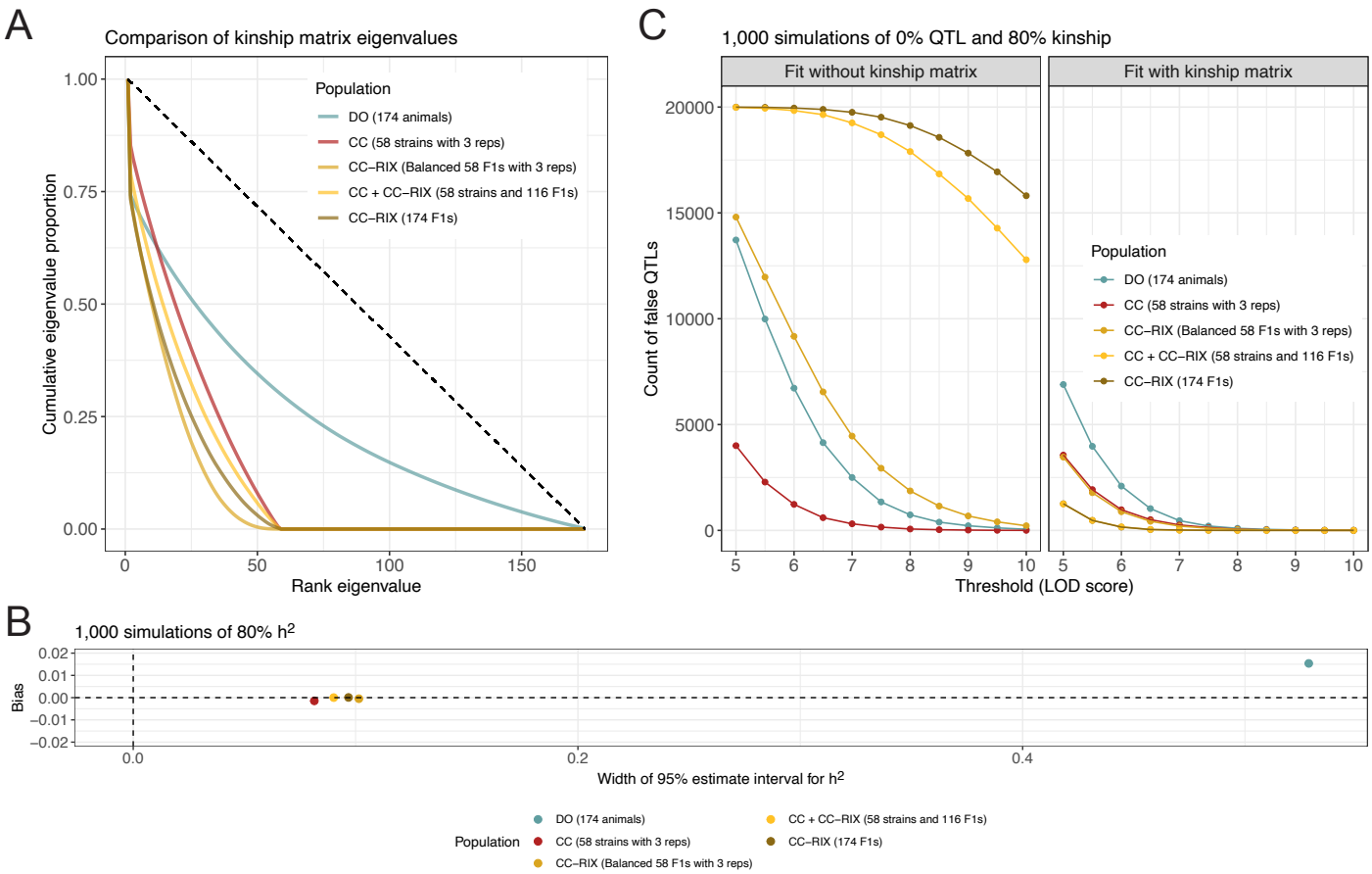

Figure S3 Population structure or unequal relatedness influence heritability estimation and QTL mapping. (A) Population structure summarized as cumulative eigenvalues (y-axis) by the rank of the eigenvalues of the kinship matrices for the CC, CC-RIX, and DO populations. Diagonal dashed line included for reference representing a population with no structure. (B) Comparing heritability estimation in terms of bias and 95% estimate interval width across the population. Summaries for each population are based on 1,000 simulations with heritability set to 80%. Horizontal and vertical dashed lines at 0 included for reference representing a perfect summary with no uncertainty or bias. (C) Population structure can result in false positive QTLs when it is not accounted for in the model. The number of false QTLs were counted from 1,000 simulations of data with no QTL and 80% polygenic effect size for all populations.

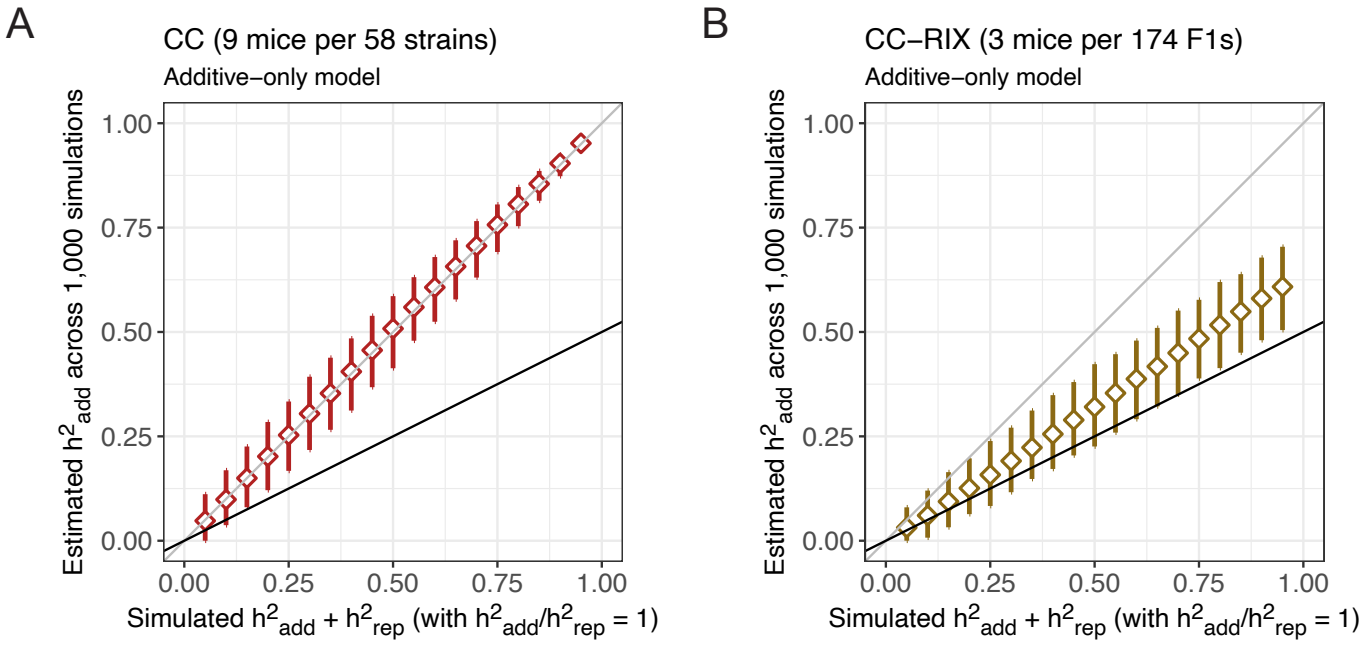

Figure S4 Performance of heritability estimation in data simulated for 522 mice from (A) CC and (B) CC-RIX populations with strain/F1 replicates when the fit model is miss-specified with only an additive component but the data are simulated with additive (h^2_{add}) and strain/F1 (h^2_{rep}) components. Diamonds represent the mean estimated heritability across 1,000 simulations from the true heritability. Vertical line segments represent middle 95% intervals across the 1,000 simulations. Black lines indicating the additive heritability component estimate is equal to the true value (y = 2x) included for reference. Gray lines indicating the additive heritability component estimate is equal to the true value of the sum of heritability components (y = x) included for reference.

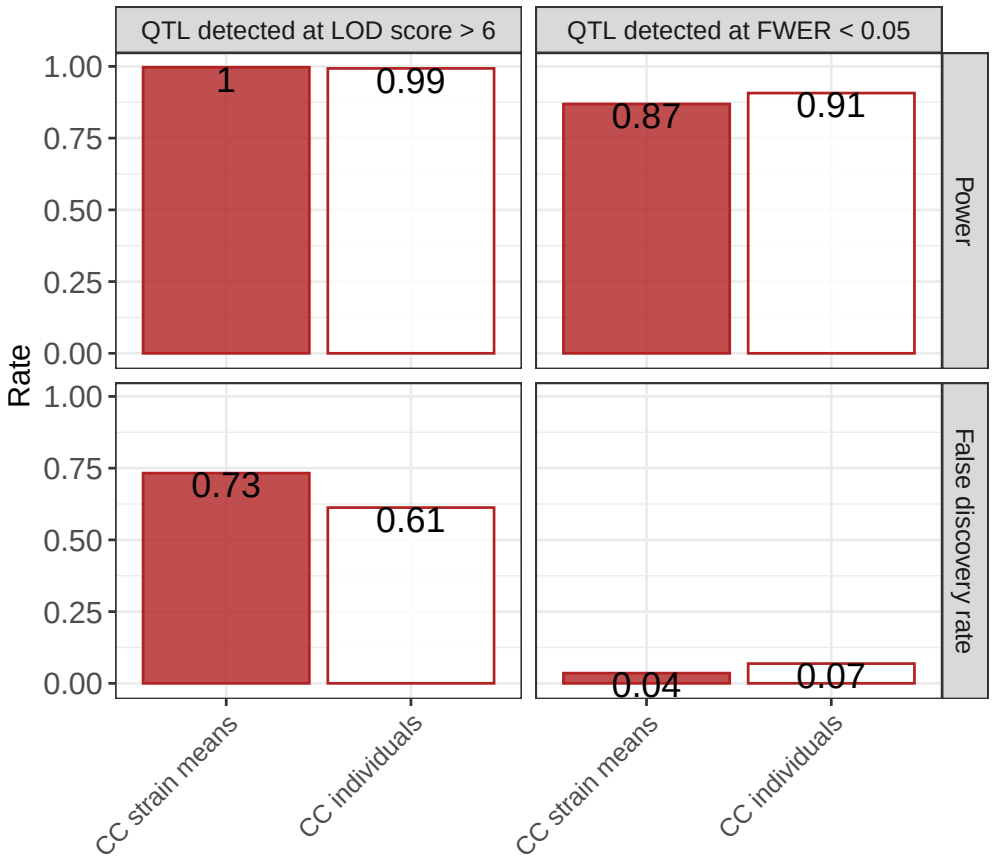

Figure S5 Comparison of power and FDR between QTL mapping based on CC strain means or the individual-level data. Rates were summarized over 1,000 simulated QTLs using significance thresholds that were lenient (LOD score > 6) and stringent (FWER < 0.05). Simulations were of 116 animals (2 animals per 58 CC strains).

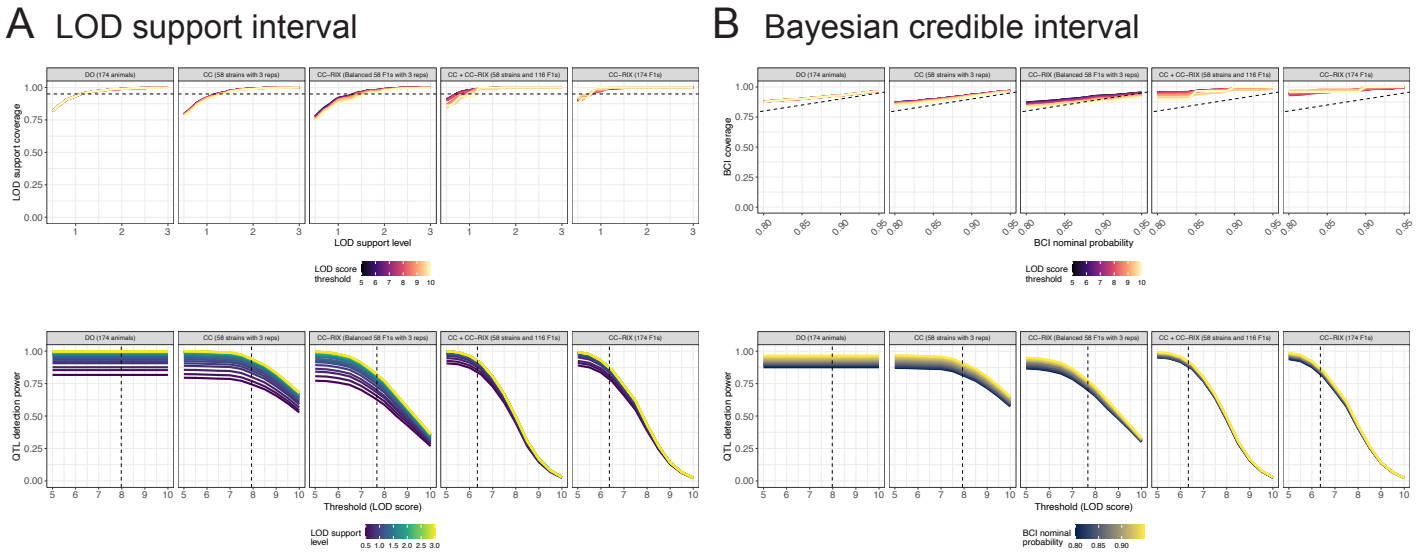

Figure S6 Performance of likelihood-based intervals for QTL location for CC, CC-RIX, and DO populations in terms of (top) QTL coverage rate and (bottom) mapping power. Likelihood-based intervals included (A) LOD support intervals and (B) Bayesian credible intervals (BCIs). Intervals are summarized over 1,000 QTLs simulated in 174 animals for each sample population. For the coverage rate of LOD support intervals, horizontal dashed lines at 80% coverage included for reference. For the coverage rate of Bayesian credible intervals, diagonal dashed lines where the nominal probability is equal to the interval coverage rate (y = x) included for reference. For mapping power, vertical dashed lines represent the 95% significance thresholds for each population. See Figure S7 for results from sampling-based intervals.

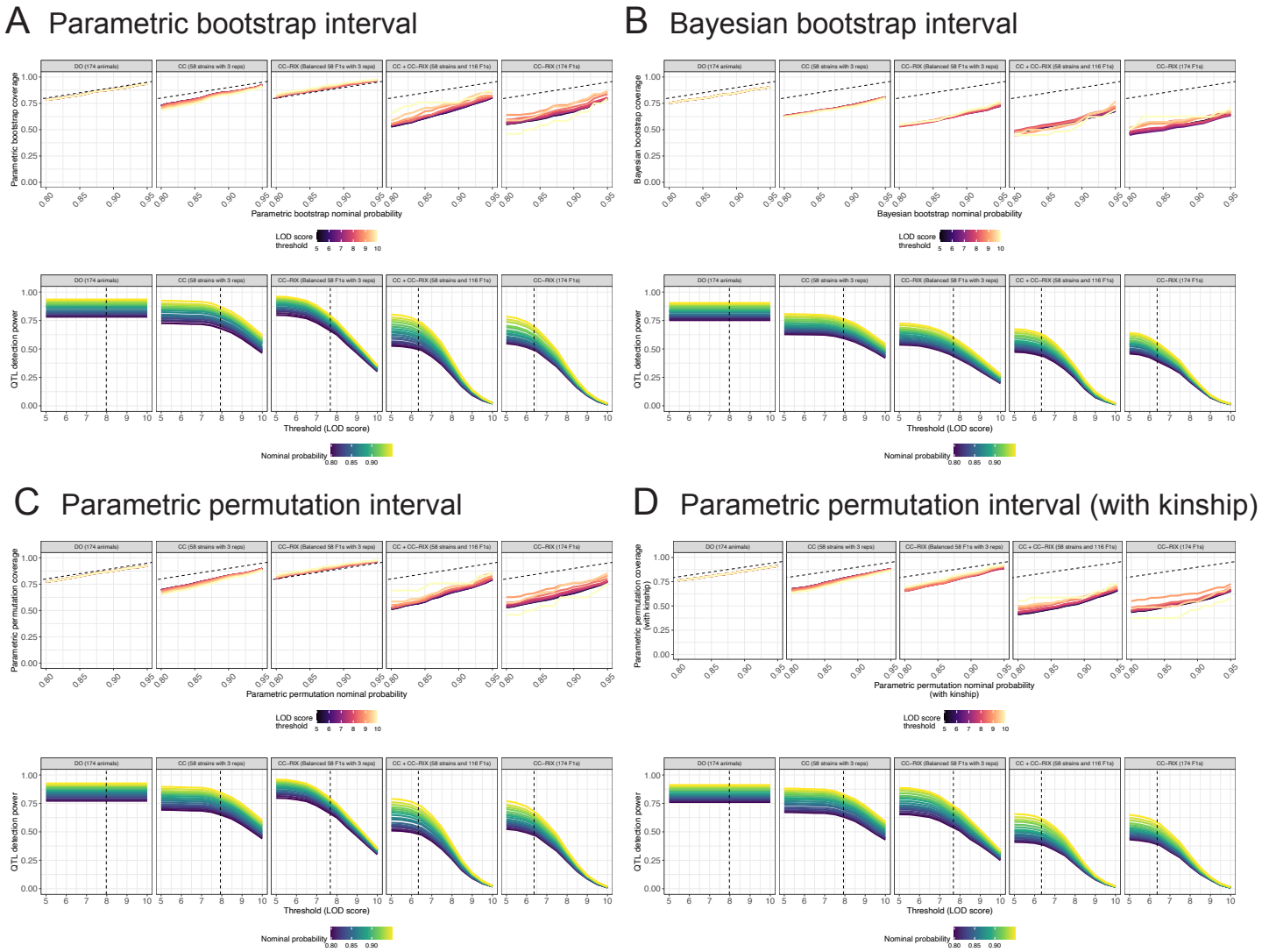

Figure S7 Performance of sampling-based intervals for QTL location for CC, CC-RIX, and DO populations in terms of (top) QTL coverage rate and (bottom) mapping power. Sampling-based intervals included (A) parametric bootstrap interval, (B) Bayesian bootstrap intervals, (C) parametric permutation intervals, and (D) parametric permutation intervals with kinship included. Intervals are summarized over 1,000 QTLs simulated in 174 animals for each sample population. For Bayesian bootstrap and parametric permutation intervals with kinship, 200 samples were generated; for the other interval types, 1,000 samples were used. For coverage rate, diagonal dashed lines where the nominal probability is equal to the interval coverage rate (y = x) included for reference. For mapping power, vertical dashed lines represent the 95% significance thresholds for each population. See Figure S6 for results from likelihood-based intervals.

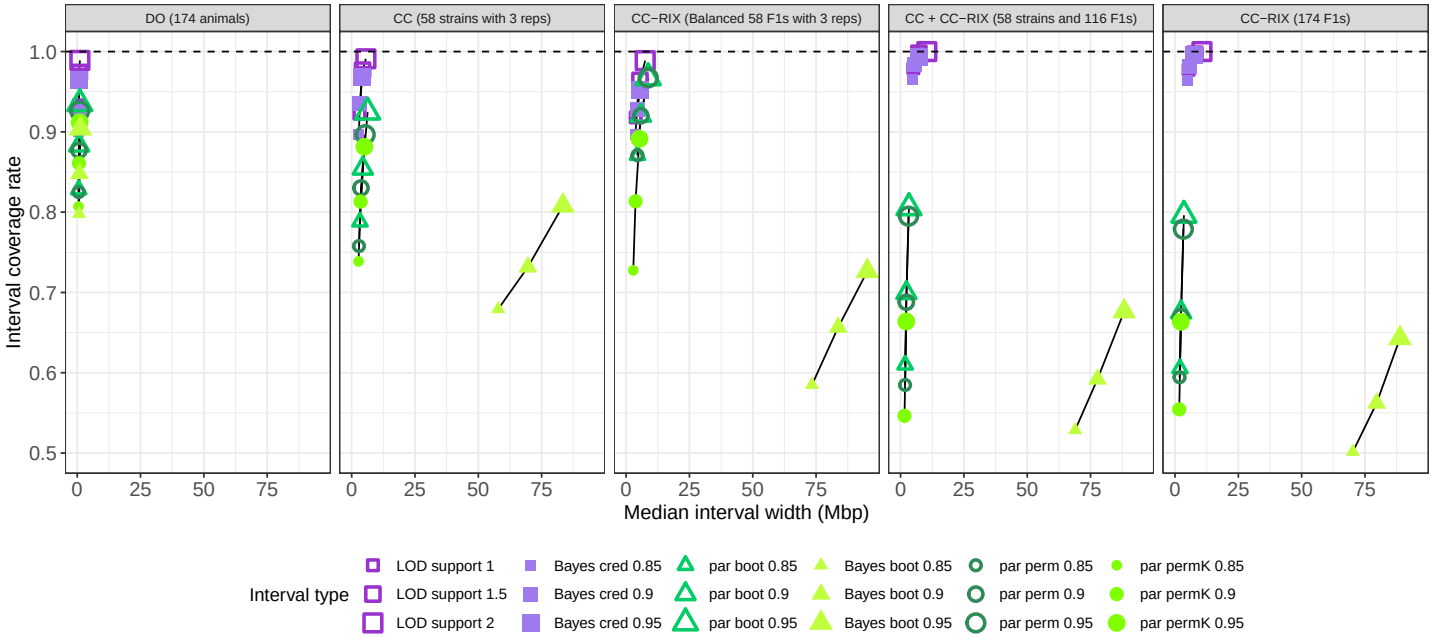

Figure S8 Performance of QTL location interval estimation across data simulated for CC, CC-RIX, and DO populations in terms of QTL coverage rate (y-axis) and interval width (x-axis). Intervals are summarized over 1,000 QTLs simulated in 174 animals for each sample population. For Bayesian bootstrap and parametric permutation intervals with kinship, 200 samples were generated; for the other sampling-based interval types, 1,000 samples were used. Purple shapes represent likelihood-based procedures and green shapes represent sampling-based procedures. Solid lines connect summaries from the same procedure but with differing support levels, with increasing support indicated by larger symbols. For Bayesian bootstrap and parametric permutation intervals with kinship, 200 samples were generated; for the other interval types, 1,000 samples were used. Horizontal dashed line representing complete coverage included for reference. The same summaries are shown in Figure 7 with Bayesian bootstrap omitted for clarity.