G3, 2021, 11(7), jkab131
DOI: 10.1093/g3journal/jkab131
Advance Access Publication Date: 23 April 2021
Investigation

Effects of kinship correction on inflation of genetic interaction statistics in commonly used mouse populations

Anna L. Tyler,[1] Baha El Kassaby,[1] Georgi Kolishovski,[1] Jake Emerson,[1] Ann E. Wells,[1] J. Matthew Mahoney,[1,2,3] and Gregory W. Carter[1,*]

[1]The Jackson Laboratory, Bar Harbor, ME 04609, USA
[2]Department of Neurological Sciences, University of Vermont, Burlington, VT 05405, USA
[3]Department of Computer Science, University of Vermont, Burlington, VT 05405, USA

*Corresponding author: The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA. Email: gregory.carter@jax.org

Abstract

It is well understood that variation in relatedness among individuals, or kinship, can lead to false genetic associations. Multiple methods have been developed to adjust for kinship while maintaining power to detect true associations. However, relatively unstudied are the effects of kinship on genetic interaction test statistics. Here, we performed a survey of kinship effects on studies of six commonly used mouse populations. We measured inflation of main effect test statistics, genetic interaction test statistics, and interaction test statistics reparametrized by the Combined Analysis of Pleiotropy and Epistasis (CAPE). We also performed linear mixed model (LMM) kinship corrections using two types of kinship matrix: an overall kinship matrix calculated from the full set of genotyped markers, and a reduced kinship matrix, which left out markers on the chromosome(s) being tested. We found that test statistic inflation varied across populations and was driven largely by linkage disequilibrium. In contrast, there was no observable inflation in the genetic interaction test statistics. CAPE statistics were inflated at a level in between that of the main effects and the interaction effects. The overall kinship matrix overcorrected the inflation of main effect statistics relative to the reduced kinship matrix. The two types of kinship matrices had similar effects on the interaction statistics and CAPE statistics, although the overall kinship matrix trended toward a more severe correction. In conclusion, we recommend using an LMM kinship correction for both main effects and genetic interactions and further recommend that the kinship matrix be calculated from a reduced set of markers in which the chromosomes being tested are omitted from the calculation. This is particularly important in populations with substantial population structure, such as recombinant inbred lines in which genomic replicates are used.

Keywords: kinship; epistasis; linear mixed model

Introduction

Variation in relatedness, or kinship, among individuals in genetic association studies can lead to artificial inflation of association statistics leading to false positives and loss of power to detect true positive associations (Devlin and Roeder 1999; Voight and Pritchard 2005; Astle *et al.* 2009). While this inflation is commonly noted in highly structured human populations, it is also often evident in laboratory crosses (Figure 1) that can also exhibit genotypic similarity. In this case, associations at a true causal quantitative trait locus (QTL) will generate association signals at all other loci with similar genotype segregation, the majority of which will likely be false positives. For multiple true positives, polygenic effects can increase apparent association when the causal loci are collinear. This inflation will likely vary with population type and study design, depending on genetic similarity of the crossed inbred strains, generations of crossing, multilocus incompatibilities, and variable recombination patterns.

There are a number of statistical strategies to remove this bulk effect of relatedness (Conomos *et al.* 2014; Sul *et al.* 2018; Yao and Ochoa 2019). One popular method of kinship correction is to estimate relatedness based on genotyped variants and model it as a random effect using linear mixed models (LMMs) as described in Kang *et al.* (2008). This method was originally developed to correct kinship effects on genetic main effects in highly structured mouse populations, such as the hybrid mouse diversity panel (Kang *et al.* 2008) or multigeneration populations from advanced intercross lines (AILs) (Cheng *et al.* 2013). The effects of kinship corrections on main effects in these types of populations are well studied and have been shown to dramatically reduce false positive rate (FPR) and increases power to detect true main effects (Kang *et al.* 2008; Cheng *et al.* 2013). Relatively unstudied, however, are the effects of population structure and relatedness on genetic interaction test statistics.

Genetic interactions, or epistasis, are an important aspect of describing complex traits. Statistical models of complex traits are improved when epistasis is taken into account (Forsberg *et al.* 2017), particularly when considering individuals at the tails of the trait distribution (Tyler *et al.* 2017). Thus, epistasis may contribute to missing heritability and poor replication of genetic associations across human populations (Mackay 2014). Kinship may influence these pairwise effects similarly to main effects.

Received: February 11, 2021. **Accepted:** March 31, 2021

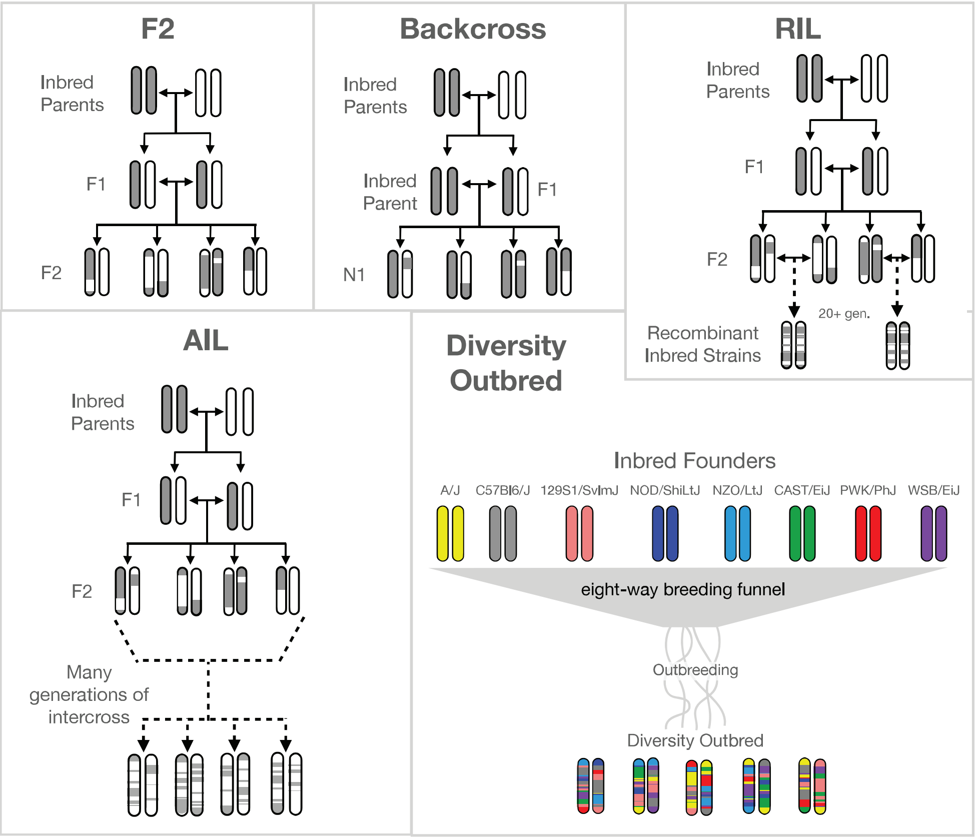

Figure 1 Breeding schemes for mouse populations used. F2—Mice from two inbred strains are bred together generating genetically identical, heterozygous mice in the F1 generation. These mice are bred together to generate mice in the F2 generation. Chromosomes in the F2 have been recombined in a single generation producing mice with unique combinations of parental chromosomes. Backcross—Mice from the F1 generation are bred with one of the parental strains creating mice in the N1 generation that have one parental chromosome, and one recombined chromosome. RIL—Mice from the recombined F1 generation are sibling mated over many generations to generate inbred mice with two identical chromosomes that are a combination of the parental chromosomes. Because each mouse is inbred, the line of identical mice can be propagated, but across the panel of inbred lines, each mouse has a unique combination of the original parental chromosomes. AIL—Rather than being inbred as with an RIL, mice in an AIL are continuously intercrossed producing genetically unique mice with high heterozygosity. Each subsequent generation of mice has more chromosomal recombinations than the prior generation. Diversity Outbred—As with the AIL, diversity outbred mice are intercrossed each generation, producing unique, highly heterozygous mice, with increasingly recombined chromosomes. In contrast to the AIL, the diversity outbred mice were started from eight inbred founders. Figures adapted from Ashbrook *et al.* (2019) and Saul *et al.* (2019).

Understanding, and appropriately adjusting for kinship when studying epistasis are important in reducing FPR while still maintaining power to detect epistatic effects, which are often weaker than main effects.

Previous work suggests that kinship inflates interaction test statistics, and that adjusting specifically for epistatatic kinship effects effectively reduces FPR in interaction test statistics and improves modeling of complex traits through genetic interaction networks (Ning *et al.* 2018). We sought to expand upon this work by surveying a range of commonly used mouse mapping populations. Although it is common to apply kinship corrections universally, the effects of these corrections are relatively unstudied across population types. Here, we investigated the effects of kinship on interaction statistics in commonly used populations that sampled a range of relatedness as well as population structure.

In addition to calculating main effects and interaction effects using linear models, we investigated the effects of kinship on genetic interaction coefficients from the Combined Analysis of Pleiotropy and Epistasis (CAPE). We previously developed CAPE to combine information across multiple traits to infer directed genetic interactions (Carter *et al.* 2012; Tyler *et al.* 2013). This type of epistasis analysis is distinct from standard single-trait epistasis analysis, in that the interactions are inferred for multiple traits simultaneously and are directional. The most recent version of CAPE published on the Comprehensive R Archive Network (CRAN) implements an LMM kinship correction, and here, we investigated whether kinship caused inflation of these statistics, and how the kinship correction would affect any observed inflation.

For this survey, we selected six mouse populations that are commonly used for identifying both main effects and interaction effects. The populations represented a range of relatedness as well as population structure. In each population, we assessed the degree of inflation of main effect test statistics, and interaction test statistics from linear models, as well as CAPE test statistics. We also implemented a kinship correction to investigate the impact of these corrections on test statistic inflation. We used the LMM method originally described in Kang *et al.* (2008). This method corrects for both cryptic relatedness and population structure simultaneously and can handle nearly arbitrary and complex genetic relationships between individuals (Sul *et al.* 2018). This is a potentially useful feature in many complicated mouse populations, such as in multigenerational outbred populations, or in experiments involving recombinant inbred lines (RILs) with genomic replicates.

We calculated two different kinship matrices for these corrections. For one, we used the full set of genotyped markers to create an overall kinship matrix. For the other, we calculated reduced kinship matrices using only markers on chromosomes not being tested for association. The overall correction has been shown to be overly stringent and can reduce power to detect main effects (Cheng *et al.* 2013). However, by calculating kinship matrices leaving out the chromosome being tested simultaneously controls FPR and retains power to detect main effects on the omitted chromosome (Cheng *et al.* 2013). This method is called leave-one-chromosome-out, or LOCO. We implemented an extension of LOCO for epistatic tests in which we calculate kinship matrices leaving out the pair of chromosomes containing the markers being tested. Of course, if both markers are on the same chromosome, this is the same as LOCO. Here, we call this extension leave-two-chromosomes-out, or LTCO. We compared the effect of these two kinship matrices across all test statistics and all populations.

Materials and methods

Data

We examined genomic inflation in previously published data sets representing commonly used mouse populations. We selected these populations to represent a range of relatedness and population structure. The populations were as follows: a reciprocal backcross (Reifsnyder *et al.* 2000), an F2 intercross (Tyler *et al.* 2016), a panel of BXD RILs (Delprato *et al.* 2015, 2017, 2018), a cohort of Diversity Outbred mice (Svenson *et al.* 2012; Tyler *et al.* 2017), and a cohort from an AIL (Hernandez Cordero *et al.* 2018). We created a sixth study population by averaging over genomic replicates in the RIL. Averaging over replicates in RILs is common practice. It not only reduces population structure but also reduces n and the power to detect effects (Keele *et al.* 2019). We refer to this population as the RIL with no replicates (RIL-NR). Breeding schemes for each population are shown in Figure 1.

We expected that the AIL, F2, and backcross populations would have negligible population structure, but may potentially harbor cryptic relatedness, where random differences in recombination led to some pairs of individuals being more highly related than other pairs of individuals. Outbred and RIL populations are more likely to have population structure that may confound genetic association tests. The Outbred population used here included multiple generations of animals, and the RIL population included genetic replicates. Each data set is described in more detail below.

Mouse populations

Advanced intercross lines

This AIL was started by crossing a large (LG/J) mouse with a small (SM/J) mouse (Cheverud *et al.* 1996). The study population used here are all males derived from the 50th filial generation (Hernandez Cordero *et al.* 2018). Mice were assessed for skeletal and muscular traits at 12 weeks of age (Hernandez Cordero *et al.* 2018). The mice were genotyped at 7187 SNPs. Here, we analyzed tibia length (Tibia) and soleus weight (Soleus) in 492 mice.

Backcross

This population was generated to investigate gene–environment interactions influencing diabetes and obesity (Reifsnyder *et al.* 2000). The diabetes-prone New Zealand Obese (NZO/HlLtJ) mouse was crossed to the diabetes-resistant Non-obese Non-diabetic (NON/ShiLtJ) mouse. The F1 generation was then backcrossed to the NON parent. The study population comprised 204 male mice genotyped at 84 microsatellite markers. For this study, we selected trygliceride level and high-density lipoprotein levels. We used cross direction (paternal grandmother abbreviated as pgm) as a covariate in all runs.

F2

This large F2 intercross was generated to investigate genetic influences on bone density traits in mice (Tyler *et al.* 2016). This population carried a fixed *lit* mutation in growth hormone-releasing hormone receptor (GHRHR), which arose naturally on the C57Bl6/J (B6) background, and was transferred to the C3H/HeJ (C3H) background. C3H mice with the *lit* mutation have the same body weight as B6 mice with the *lit* mutation but have higher bone density. The purpose of this cross was to identify genetic factors that increase bone density in the absences of GHRHR. We used 1095 female mice from this cross. They were genotyped at 100 microsatellite markers. We analyzed percent body fat (pctFat) and trabecular bone thickness (Tb.Th) here.

Outbred

We used a cohort of Diversity Outbred mice (Svenson *et al.* 2012), which were derived from eight founder strains: 129S1/SvImJ (129), A/J, CAST/EiJ (CAST), NOD/ShiLtJ (NOD), NZO/HlLtJ (NZO), PWK/PhJ (PWK), and WSB/EiJ (WSB). The CAST, PWK, and WSB strains were recently inbred from wild mouse strains, whereas the other five strains were inbred mostly from pet fancy mice with limited genetic diversity (Yang *et al.* 2007). Across all eight strains, there are roughly 45 million SNPs, and because DO mice are outbred, each carries a unique subset of these SNPs. The systematic mating scheme was designed to limit population structure and relatedness. The DO population we used here included 446 individuals, both male and female. We used only the mice that were fed on a chow diet, eliminating those on a high-fat diet. We used sex as a covariate in all runs. We analyzed the change in blood glucose between 6 and 19 weeks of age (change.urine.glucose) and blood glucose levels at 19 weeks of age (urine.glucose2) in this study.

Recombinant inbred lines

The RILs we analyzed here were from the BxD panel of RILs. RILs are generated by crossing two parental strains, breeding the progeny for some number of generations to produce recombinant chromosomes, and then inbreeding to generate stable, inbred genotypes. The result is a panel of inbred mice each with a unique combination of genotypes from the parental strains. BxD were generated from an initial cross between the C57Bl/6J (B) mouse and the DBA/2J (D) mouse. We downloaded data from the MousePhenomeDatabase (Grubb *et al.* 2014) on August 5, 2020.

The data we analyzed were from an experiment investigating the genetics of hippocampal anatomy and spatial learning (Delprato *et al.* 2015, 2017, 2018). The data set is called Crusio1. We downloaded all traits related to body weight, radial maze performance, and histopathology.

The BxD panel has been genotyped at 7124 markers across the genome. The genotypes are available from GeneNetwork (Sloan *et al.* 2016). We analyzed time to complete the radial maze on the first day of training (task_time_d1) and the number of radial arms entered on day 5 of training (num_arms_d5) in 452 females.

Recombinant inbred lines, no replicates

This test used the same RIL data set as described above, but we averaged over individuals of the same strain resulting in

55 individuals. Averaging over replicates in a single strain is common practice. This practice not only reduces structure in the mapping population but also reduces power to detect effects (Keele *et al.* 2019). Here, we examined how averaging across replicates in a strain affected test statistic inflation. We used only females in this analysis to completely eliminate any duplicated genomes.

Trait selection

CAPE combines information across multiple traits and requires at least two traits as input. It has been observed previously that body weight and size traits are significantly correlated with the proportion of NZO genotype in an individual (P. Simicek, personal communication). Two of our populations here, the backcross and the outbred populations, include NZO genomes, and using body weight traits in these populations could lead to increased test statistic inflation due to high levels of polygenicity. To reduce this effect, we selected traits from each population that minimized the correlation with the first principal component of the kinship matrix (Supplementary Figure S1 and Table S1).

Kinship matrix calculation

We use the R package qtl2 (Broman *et al.* 2019) to calculate the kinship matrix as described in Kang *et al.* (2008). This method calculates a similarity matrix based on measured genotypes. This matrix has been shown to correct confounding population structure effectively and is guaranteed to be positive semidefinite. The kinship matrix is calculated as follows:

$$K = \frac{G \times G^T}{n},$$

where G is the genotype matrix, and n is the number of genotyped markers. For calculating main effects, we use the LOCO method (Cheng *et al.* 2013), in which the markers on the chromosome being tested are left out of the kinship matrix calculation. LOCO has been shown to reduce the rate of false negatives relative to use of the overall kinship matrix (Cheng *et al.* 2013; Gonzales *et al.* 2018). For each chromosome, we calculated:

$$K_C = \frac{G_C \times G_C^T}{n},$$

where G_C is the genotype matrix with all markers on chromosome C removed. For the pairwise tests, we used the natural extension of LOCO, which we called LTCO. To calculate the kinship matrix for a pairwise test, we left out the two chromosomes containing the two markers being tested. If both markers were on the same chromosome, we left out only that one chromosome.

F_{ST}

To gauge the level of population structure in each population, we calculated the fixation index (F_{ST}) using the sum of the heterozygosity across all loci π (Hahn 2019). In the following equation, π_T is the heterozygosity across all populations and π_S is the average heterozygosity across the subpopulations (Hahn 2019).

$$F_{ST} = \frac{\pi_T - \pi_S}{\pi_T}.$$

An F_{ST} of 0 indicates that the population is interbreeding freely, and a value of 1 indicates that subpopulations within the population are genetically isolated. Here, F_{ST} estimated how structured each mouse population was.

To do this, we converted the kinship matrix for each population to a weighted network using the R package igraph (Csardi and Nepusz 2006). We used the fast-greedy clustering algorithm (Clauset *et al.* 2004) in igraph to define subpopulations, which we then used to calculate F_{ST}.

LMM correction

To account for population structure in our association tests, we used an LMM correction as described in Lippert *et al.* (2011) and Kang *et al.* (2008). Briefly, population corrections account for polygenic effects on the phenotype that are not attributable to the test marker, which cause the assumption of independent prediction errors to fail. To account for correlated errors, Kang *et al.* proposed a mixed-effects model where the residual errors are not independent, but correlated according to a multivariate Gaussian distribution whose covariance matrix is given by a linear combination of the identity matrix (independent random noise) and a kinship matrix, K, which is simply the variance-covariance matrix of the genotypes among individuals. Fitting this model requires identifying the maximum likelihood parameters for the genetic (fixed) effects and the two mixing parameters defining the correlated residual errors. As shown by Lippert *et al.* (2011), this model can be fit rapidly by first factoring K into its spectral decomposition and adjusting the genotypes and phenotypes to align with the residual error structure.

The mathematical form of the model allows the fixed effects and the genetic variance to be solved for explicitly as a function of a mixing parameter, which can be optimized using a one-dimensional grid search. We have reimplemented this procedure within CAPE for use with mouse model populations using the code from the R/qtl2 implementation (Broman *et al.* 2019).

Test statistics from linear models

After adjusting for kinship effects, we used single-locus marker regression and pairwise marker regression to derive test statistics in each population. For the single-locus regression, we fit the following model:

$$U_i^j = \beta_0^j + \sum_{c=1}^{n_c} x_{c,i}\beta_c^j + x_{1,i}\beta_1^j + \epsilon_i^j,$$

where U corresponds to traits, and ϵ is an error term. The index i runs from 1 to the number of individuals, and j runs from 1 to the number of traits. x_i is the probability of the presence of the alternate allele for individual i at locus j. We calculated P-values for each test statistic analytically using a t distribution with n—1 degrees of freedom, where n was the number of individuals in the population. We collected main effect test statistics for all traits in each data set.

For the pairwise marker scans, we limited our analysis to two traits. As described below, CAPE requires at least two traits. However, CAPE and pairwise tests in general are computationally intensive, and our ability to run many traits was limited. We fit linear models for each pair of markers and each of the two selected traits as follows:

$$U_i^j = \beta_0^j + \underbrace{\sum_{c=1}^{n_c} x_{c,i}\beta_c^j}_{\text{covariates}} + \underbrace{x_{1,i}\beta_1^j + x_{2,i}\beta_2^j}_{\text{main effects}} + \underbrace{x_{1,i}x_{2,i}\beta_{12}^j}_{\text{interaction}} + \epsilon_i^j.$$

Again, U corresponds to traits, and ϵ is an error term. The index i runs from 1 to the number of individuals, and j runs from 1 to the number of traits. x_i is the probability of the presence of the alternate allele for individual i at locus j. For the pairwise tests, we calculated empirical P-values from permutations. We combined test statistics from individual permutations to generate a null distribution with 1.5 million randomized values (Carter *et al.* 2012).

Combined Analysis of Pleiotropy and Epistasis

Starting with the pairwise linear regression above, we ran the CAPE (Carter *et al.* 2012; Tyler *et al.* 2013). CAPE reparametrizes β coefficients from pairwise linear regressions to infer directed influence coefficients between genetic markers. The reparametrization combines information across multiple traits thereby identifying interactions that are consistent across all traits simultaneously. Combining information across traits also allows inference of the direction of the interaction (Carter *et al.* 2012; Tyler *et al.* 2013).

The β coefficients from the linear models are redefined in terms of two new δ terms, which describe how each marker either enhances or suppresses the activity of the other marker:

$$\begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} = \begin{bmatrix} \beta_1^1 & \beta_2^1 \\ \beta_1^2 & \beta_2^2 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \beta_{12}^1 \\ \beta_{12}^2 \end{bmatrix}.$$

We then translated the δ terms into marker-to-marker influence terms:

$$\delta_1 = m_{12}(1+\delta_2),\ \delta_2 = m_{21}(1+\delta_1).$$

Since matrix inversion can lead to large values with larger standard errors, we performed standard error analysis on the regression parameters, and propagated the errors using a second-order Taylor expansion (Bevington 1994; Carter *et al.* 2012). To calculate P-values for the directed influence coefficients, we performed permutation testing.

Evaluation of inflation

We ran CAPE on each population using each of the population corrections: none, LMM-overall, or LMM-LOCO. For each run, we collected the main effect statistics and interaction effect statistics from the linear models, as well as the CAPE statistics. To estimate the variation in test statistic distributions across sampled populations, we performed Monte-Carlo cross-validation (Xu and Liang 2001) by sampling 80% of the individuals over 10 trials.

In each trial, we assessed the inflation of each set of test statistics using λ (Devlin and Roeder 1999). This inflation factor is the ratio of the median test statistic over the mean of the theoretical distribution. Here, we calculated the mean of the chi-square quantiles of $1-p$ over the theoretical mean of the null, uniform P-value distribution with one degree of freedom (0.456).

Data availability

All data used in this study and the code used to analyze it are available as part of a reproducible workflow on Figshare. The workflow is called Epistasis_and_Kinship_in_Mouse_Populations and can be found here: https://figshare.com/articles/journal_contribution/Epistasis_and_Kinship_in_Mouse_Populations/13863101.

CAPE is available at CRAN and on Github at https://github.com/TheJacksonLaboratory/cape.

Supplementary File S1 contains descriptions of each of the supplemental files. Supplementary Figure S1 shows box plots of the correlations between traits and the first PC of the kinship matrix for each population. Supplementary Figure S2 is identical to Fig. 3, but with the addition of the lambda values for the subsampled F2 population. Supplementary Figure S3 shows the correlations between CAPE and main effect lambda statistics. Supplementary Table S1 contains Pearson correlations between traits and the first PC of the kinship matrix.

Results

Population structure and relatedness varied across populations

We observed varying degrees of relatedness across the populations (Figure 2). The heat maps in Figure 2 show how each population was clustered into subpopulations and the relatedness within and among subpopulations. The AIL and F2 populations had negligible structure with no discernible differences in heterozygosity across subpopulations. The Outbred mice were drawn from multiple generations of DO mice, which created subpopulations with slightly higher relatedness than the overall average. The Backcross and RIL had the most substantial structure, with F_{ST} values more similar to human populations.

Independent of the population structure, the populations also had varying degrees of relatedness. On average, the Outbred mice were related to each other at a level equivalent to first cousins, which is by design (Svenson *et al.* 2012), whereas the AIL mice were slightly more related to each other than siblings. The other three populations were all siblings on average, but had differential variation around that mean, with the F2 having a very narrow distribution of relatedness and the Backcross having a wider distribution (Figure 2).

Main effect test statistic inflation varied widely across populations

Before running CAPE, we investigated overall trends in test statistic inflation by scanning all traits for main effects using marker regression. This revealed wide variation in test statistic inflation by population (Figure 3). Across all traits, the AIL, Outbred, and RIL-NR populations showed very little inflation. In contrast, the RIL, F2, and Backcross populations showed substantial inflation across most or all traits when no kinship correction was applied (Figure 3, left-most group). Inflation in the RIL population was corrected by an LOCO kinship correction (Figure 3, middle group). The overall kinship correction eliminated inflation in all populations (Figure 3, right-most group).

Main effect inflation was correlated with linkage disequilibrium

Linkage disequilibrium (LD) influences test statistic inflation because a single causal SNP within an LD block can inflate the test statistics of all SNPs linked to it. If there are relatively few recombinations in the population, such as in an F2 or backcross, large portions of the genome may be significantly associated with a trait due to linkage alone.

To investigate whether LD may be related to the inflation of test statistics in the populations used here, we calculated

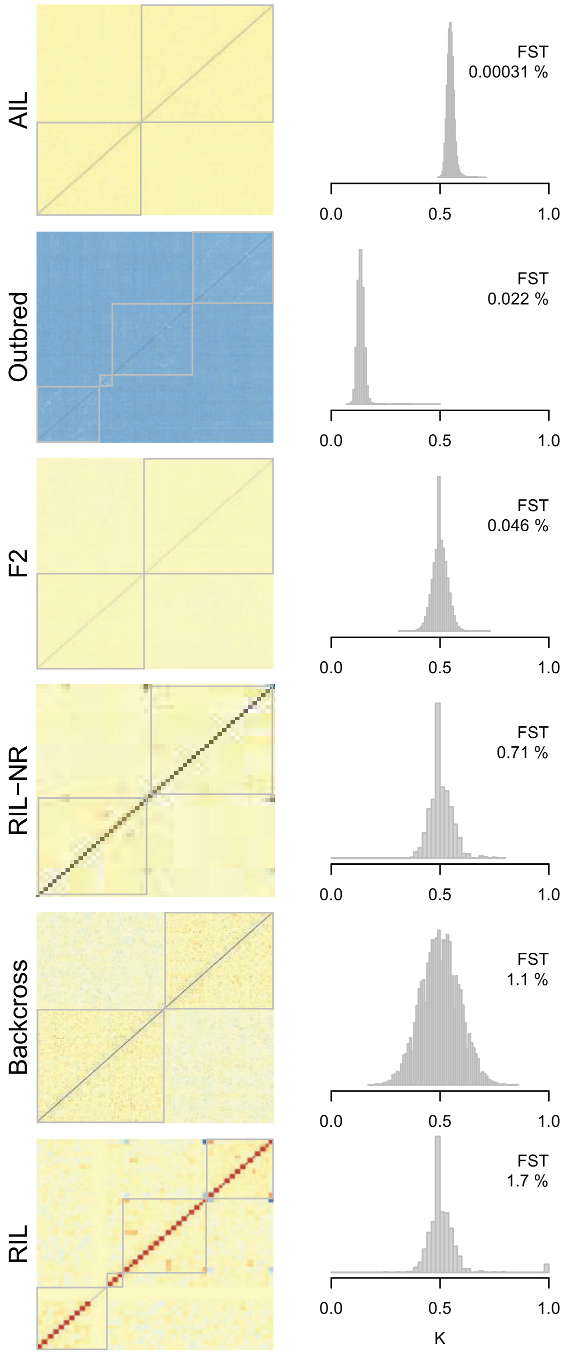

Figure 2 Population structure and relatedness distributions across populations. Each panel shows the population structure of a population as a heat map on the left side of the panel. The heat maps show the overall kinship matrix for each population clustered into subpopulations. Cool colors indicate less relatedness, and warm colors indicate more relatedness. The gray lines indicate the boundaries of subpopulations. The histograms to the right of each heat map show the distribution of relatedness from the upper triangle of the kinship matrix. Average relatedness varies from the level of cousins in the Outbred animals to slightly higher than the level of siblings in the AIL animals. The F_{ST} value for each population is shown in the upper right-hand corner of each histogram and was calculated as described in the methods. Panels are in order of increasing F_{ST}.

pairwise Pearson correlations (r) between markers on the same chromosome across all chromosomes and all populations. These distributions are shown in the inset in Figure 3. The two populations with the highest test statistic inflation, the F2 and Backcross populations, also had the highest average LD.

However, although the F2 had lower LD than the backcross, it had substantially greater inflation of test statistics. The F2 also had many more individuals than the backcross, and thus greater power to detect effects. This increase in power combined with high LD could lead to the high levels of inflation seen in the F2. To test this, we subsampled the F2 to the same number of individuals in the backcross and recalculated λ. Reducing n in the F2 also reduced inflation to similar levels seen in the backcross (Supplementary Figure S2).

Kinship corrections reduced inflation differentially across populations

Figure 4A shows a more detailed view of test statistic inflation in the main effect statistics for each population. Each panel shows quantile–quantile (QQ) plots for the $-log_{10}(p)$ for two traits against the theoretical null P-values. The more the points rise above the line $y = x$, the stronger the inflation factor λ. In the absence of a kinship correction, the F2 and RIL showed strong inflation, the AIL and RIL without replicates showed moderate inflation, and the backcross and Outbred populations showed very minor inflation if any at all (Figure 4A).

Numeric values are shown in the legends of Figure 4. The RIL (λ =2.7) was the most affected by inflation, while traits in the Outbred population had mild deflation ($\lambda = 0.82$).

The overall kinship correction had a strong effect on inflation across all populations (purple dots in Figure 4A). The LOCO correction had varied effects (green dots in Figure 4A). It provided strong control of inflation in the RIL, but had no effect in the Backcross, F2, or AIL populations.

Interaction coefficients were largely unaffected by genomic inflation

The P-values associated with interaction coefficients were almost completely unaffected by inflation (Figure 4B). The only population where inflation appeared to affect the interaction statistics was the RIL population ($\lambda = 1.1$). This inflation was reduced by both the LTCO and overall kinship corrections.

The LTCO correction appeared to slightly improve power to detect interaction effects in the Outbred population, although this was not evident in the λ values ($\lambda_{none} = 1$, vs λ_{ltco} =1).

CAPE coefficients were intermediately affected by inflation

The CAPE coefficients influenced by inflation at a level in between that of the main effect statistics and the pairwise statistics (Figure 4C). The bulk of the inflation was seen in the RIL ($\lambda = 1.7$), F2 ($\lambda = 1.4$), and backcross populations ($\lambda = 1.1$). The overall and LTCO kinship corrections had remarkably similar effects across all populations.

Discussion

In this study, we examined inflation of main effect and genetic interaction statistics in six mouse mapping populations. We also investigated the effect of kinship corrections on this inflation.

We found large variation in test statistic inflation across populations and across traits. Across populations, the primary driving factors of inflation were LD and population size. Populations with high LD, like the F2 and backcross, had the highest inflation, which is consistent with previous observations (Yang *et al.* 2011). In effect, a high degree of LD makes the distribution of alleles less

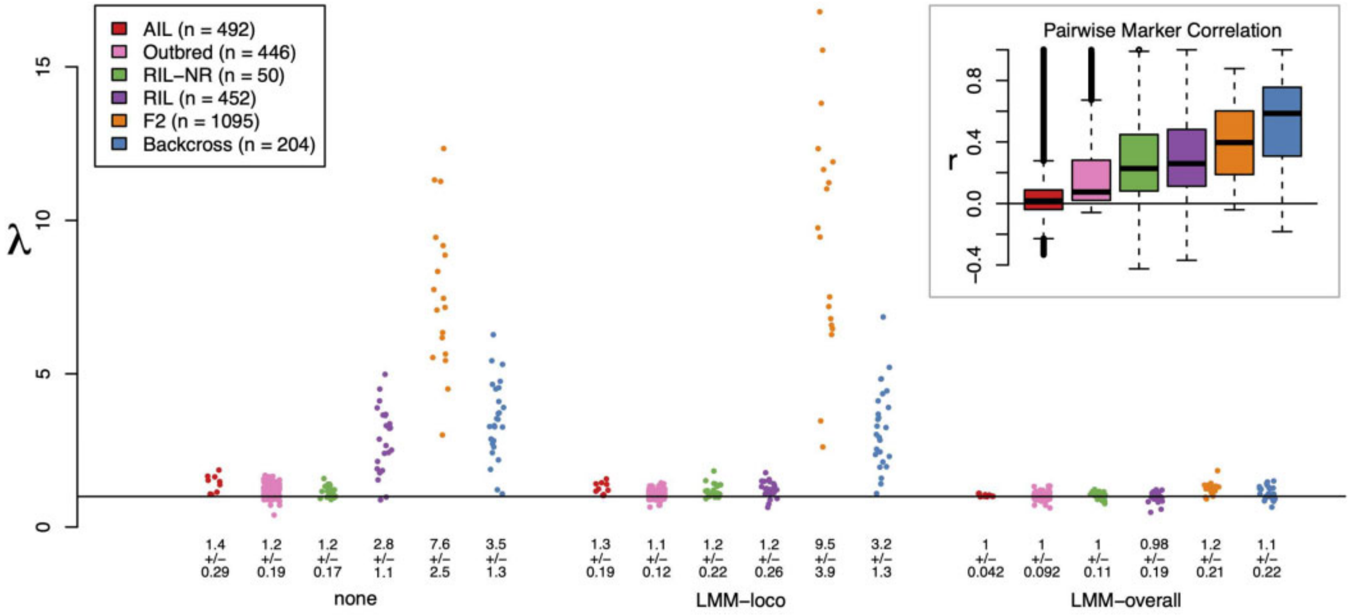

Figure 3 Inflation of test statistics for main effects. Each group of dots shows inflation of main effect statistics across all populations for one of the kinship correction types (none, LMM-loco, or LMM-overall). Each dot represents one trait. The populations are differentiated by color and are shown in order of increasing LD. The legend shows the correspondence between color and population, as well as the number of individuals in each study. The horizontal line shows $\lambda = 1$, which indicates no inflation. Numbers below each set of dots indicate the mean and standard deviation of λ for each group. The inset in the top right-hand side of the plot shows the pairwise correlation between markers on the same chromosome for each population, which is a stand-in for LD. The color of each box identifies which population the data come from. The horizontal line in the box plot shows $r = 0$. The F2 and Backcross populations, which have the highest LD, also have the highest test statistic inflation. The extreme inflation seen in the F2 population is likely due to a combination of high LD and large n.

random across individuals. Two mice that share any randomly selected allele are more likely to share a large number of alleles.

Between those populations with the highest LD, the number of individuals in the population had a large effect on inflation. Thus, high power to detect effects combined with high LD creates hugely inflated test statistics. There was also wide variation in inflation across different traits. We hypothesize that polygenicity may be the primary factor in the variation in inflation across traits within a single population. All else being equal, there will be a preponderance of small *P*-values for traits with multiple true positive loci.

Differences in LD cannot explain the difference in inflation between the RIL with replicates, which had substantial inflation, and the RIL without replicates (RIL-NR), which did not. It has been shown previously that including genetic replicates increases power to detect genetic effects (Keele *et al.* 2019). Increase in power alone potentially increases the prevalence of small *P*-values; however, genetic relatedness also increases FPR when strain effects are large relative to individual error (Keele *et al.* 2019). Taken together, these results suggest that including genetic replicates in an RIL study increases power to detect effects, but that an LOCO kinship correction should be done to counteract the increase in FPR caused by the replicates. Here, the LOCO kinship correction substantially reduced inflation in the RIL population without the reduction in power seen with the overall kinship correction (see Figures 3 and 4A).

The differences of effects between the overall and reduced kinship matrices for the main effects illustrate a couple important points about these two corrections. First, the overall kinship correction reduces power to detect true effects (Cheng *et al.* 2013). Indeed, we saw complete elimination of inflation across all populations with this correction. Second, the comparison between the LOCO and overall corrections suggests that the inflation seen in the RIL was primarily due to population structure. The substantial inflation of main effect test statistics in the RIL was reduced by the LOCO correction. However, the LOCO correction did not reduce inflation in the F2 or backcross. These populations had very little structure, and inflation was likely due primarily to LD and polygenicity.

That the overall kinship correction erased all inflation shows how this severe correction can eliminate power to detect true effects. The LOCO correction, however, retains power to detect true effects, while still correcting for relatedness. It should be noted that treating the F2 and Backcross populations as genome-wide association study (GWAS) mapping populations is not really a fair representation, since in practice the markers in these populations would not be treated as independent measurements. However, this exercise illustrates important, albeit dramatic, aspects of test statistic inflation, and how kinship corrections affect test statistics in different situations.

The interaction β coefficients did not show any inflation in any population except possibly in the RIL, despite these populations being well powered to detect epistasis (Stich and Gebhardt 2011). The effects of both type of kinship correction were minimal; however, there may have been some minor improvement of power from both corrections in the Outbred population.

This complete lack of inflation is in contrast to previous studies that have identified pairwise test statistic inflation due to population structure (Stich and Gebhardt 2011; Ning *et al.* 2018). There are a number of possible explanations for this discrepancy. Because CAPE relies on partial pleiotropy to dissect directed epistatic interactions, and because we cannot exhaustively test all pairs of markers in modern genomic studies, we select subsets of markers with large main effects for pairwise testing. This selection may reduce the size of the interaction coefficients we measure because the majority of trait variance is explained by

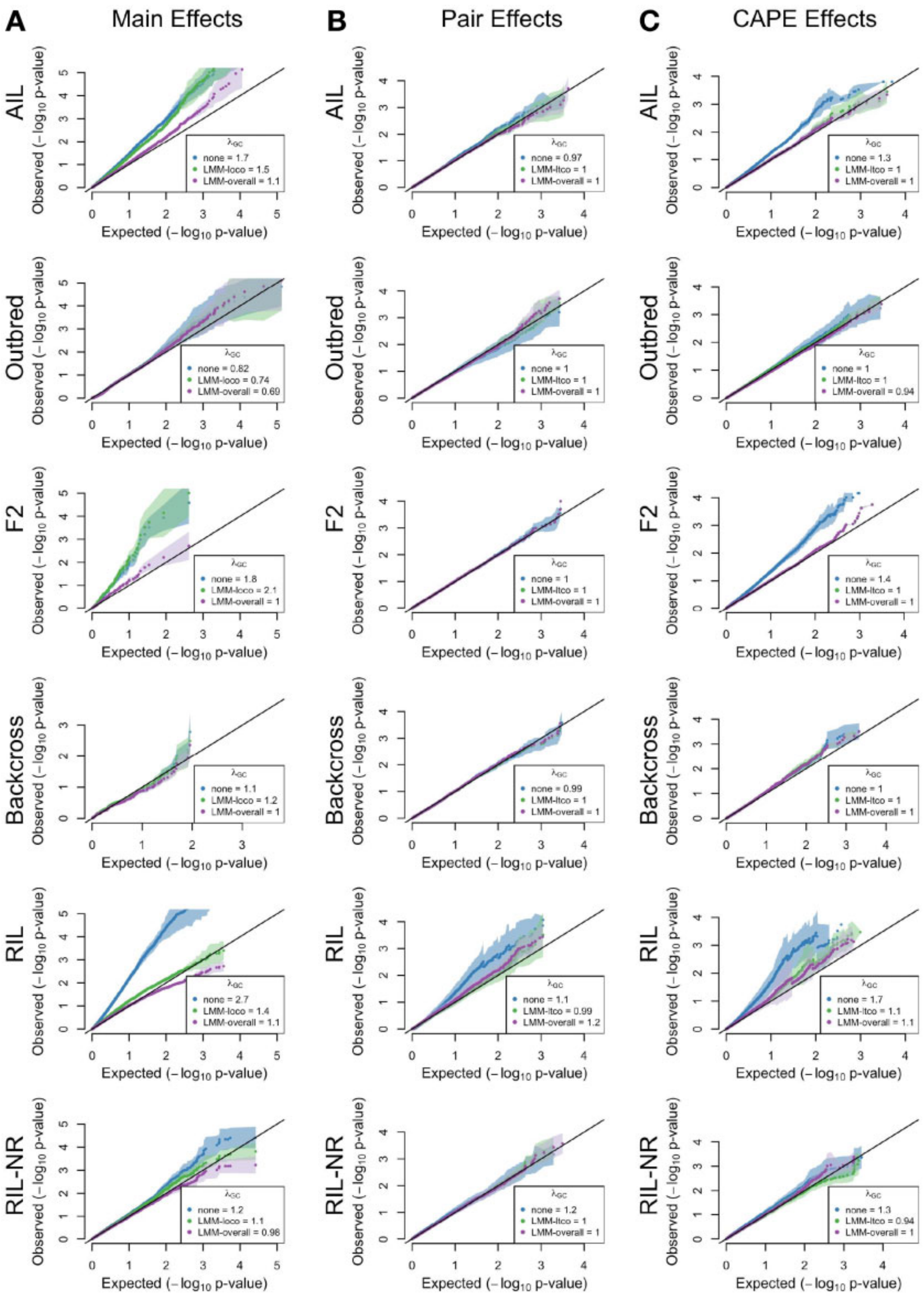

Figure 4 QQ plots for all test statistics. Each panel shows the QQ plots for one set of statistics across all populations and all correction types. Each row holds the results for a single population. Each column shows one test statistic: (A) QQ plots for main effects. (B) QQ plots for the pairwise test statistics. (C) QQ plots for CAPE statistics. Correction types (none, LMM-LOCO/LTCO, or LMM-overall) are shown in different colors. The *x*-axis in each plot shows the theoretical quantiles of the null *P*-value distribution and the *y*-axis shows the observed quantiles. Dots show the mean *P*-value distribution across 10 rounds of Monte-Carlo cross-validation, and transparent polygons show the standard deviation. The black line in each plot shows $y = x$. The legends show the λ values for each set of statistics.

marginal effects (Xu and Jia 2007; Phillips 2008; Stich and Gebhardt 2011).

Furthermore, in all but the Backcross, we tested for additive-by-additive epistasis (Cockerham 1954; Mackay 2014), which may further reduce our power to detect epistasis depending on trait heritability and allele frequencies (Stich and Gebhardt 2011).

The RIL was the only population for which there was apparent inflation in pairwise test statistics. In this population, both LMM paradigms corrected the inflation. This result is concordant with previous findings that LMM kinship corrections reduce inflation in pairwise test statistics (Stich and Gebhardt 2011; Ning *et al.* 2018).

In contrast to the interaction coefficients from pairwise linear models, CAPE interaction coefficients did show inflation in some populations. We saw the most inflation in the F2, RIL, and AIL populations. Lambda values were intermediate between those seen for the main effect statistics and the interaction statistics, which we expect given that CAPE interaction coefficients are nonlinear combinations of main effect statistics and interaction statistics across multiple traits. We therefore attribute inflation in these CAPE coefficients to propagation of main effect inflation. Indeed, the lambda values of the main effect statistics and CAPE interaction coefficients were positively correlated (Supplementary Figure S3). When there was inflation of CAPE coefficients, both corrections controlled the inflation well. The similarity in effects of the two corrections was somewhat surprising. We predicted that as with LOCO, the LTCO correction would have been less stringent than the overall correction, but this was not what we observed. Extrapolating from the main effect results, test statistic in the RIL should be most subject to inflation derived from kinship. In this population, both kinship matrices controlled inflation well, but the overall correction did trend toward the more severe correction. Although more work needs to be done, these results suggest that using the LTCO kinship matrix for interaction effects may maintain power to detect effects better than the overall matrix.

We conclude that although these association mapping populations are created in such a way as to minimize population structure, cryptic relatedness and population structure may still increase FPR and decrease power to detect both main effects and genetic interactions. This is particularly true in populations with unusual relatedness patterns, such as RILs with genomic replicates. These findings highlight the importance of examining the kinship matrix of a study population as well as *P*-value distributions across all traits. In conjunction, relatedness and the degree to which low *P*-values are enriched can reveal important features of the population and traits, including the degree to which LD and polygenicity may be influencing significance testing. In all populations, but particularly in those with greater structure, applying a kinship correction reduces FPR and increases power to detect true effects. However, our empirical analysis also found that pairwise interactions were generally avoided inflation in most study designs. The one potential exception was recombinant inbred designs (RIL), suggesting that traits with an expectation of epistasis may be more effectively analyzed with the other experimental approaches. For integrative analyses such as CAPE, an intermediate level of inflation may be best avoided with balanced outbred populations.

We recommend applying the reduced kinship matrix in which the chromosomes containing the tested markers are left out. These kinship matrices reduce FPR related to population structure with minimal effect on power. The kinship matrices calculated from the full genome reduced power to detect effects, particularly in the RIL. Simulations were beyond the scope of this project, but could potentially further delineate guidelines for when kinship corrections are necessary, and which types of kinship matrices to use. Such simulations should take LD, polygenicity, and multiple types of population structure into account.

Acknowledgments

The authors thank the reviewers of this manuscript for their careful reading and their thoughtful and helpful comments. Their work helped improve and clarify this manuscript.

Funding

This work was funded by the National Institutes of Health, National Institute of General Medical Sciences grant R01 GM115518.

Conflicts of interest

None declared.

Literature cited

Ashbrook DG, Arends D, Prins P, Mulligan MK, Roy S, *et al.* 2019. The expanded BXD family of mice: a cohort for experimental systems genetics and precision medicine. bioRxiv. Doi:10.1101/672097.

Astle W, Balding DJ. 2009. Population structure and cryptic relatedness in genetic association studies. Statist Sci. 24:451–471.

Bevington PR. 1994. Data Reduction and Error Analysis for the Physical Sciences, ISE. New York: McGraw-Hill.

Broman KW, Gatti DM, Simecek P, Furlotte NA, Prins P, *et al.* 2019. R/qtl2: software for mapping quantitative trait loci with high-dimensional data and multiparent populations. Genetics. 211: 495–502.

Carter GW, Hays M, Sherman A, Galitski T. 2012. Use of pleiotropy to model genetic interactions in a population. PLoS Genet. 8: e1003010.

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 (Bethesda). 3:1861–1867.

Cheverud JM, Routman EJ, Duarte F, van Swinderen B, Cothran K, *et al.* 1996. Quantitative trait loci for murine growth. Genetics. 142:1305–1319.

Clauset A, Newman ME, Moore C. 2004. Finding community structure in very large networks. Phys Rev E. 70:066111.

Cockerham CC. 1954. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics. 39:859–882.

Conomos MP, Miller M, Thornton T. 2014. Robust population structure inference and correction in the presence of known or cryptic relatedness. bioRxiv. doi:10.1101/008276.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. InterJ Complex Syst. 1695:1-9.

Delprato A, Algéo M-P, Bonheur B, Bubier JA, Lu L, *et al.* 2017. QTL and systems genetics analysis of mouse grooming and behavioral responses to novelty in an open field. Genes Brain Behav. 16: 790–799.

Delprato A, Bonheur B, Algéo M-P, Murillo A, Dhawan E, *et al.* 2018. A quantitative trait locus on chromosome 1 modulates intermale aggression in mice. Genes Brain Behav. 17:e12469.

Delprato A, Bonheur B, Algéo M-P, Rosay P, Lu L, *et al.* 2015. Systems genetic analysis of hippocampal neuroanatomy and spatial learning in mice. Genes Brain Behav. 14:591–606.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics. 55:997–1004.

Forsberg SKG, Bloom JS, Sadhu MJ, Kruglyak L, Carlborg O. 2017. Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. Nat Genet. 139: 1455.

Gonzales NM, Seo J, Cordero AIH, Pierre CLS, Gregory JS, *et al.* 2018. Genome wide association analysis in a mouse advanced intercross line. Nat Commun. 9:12.

Grubb SC, Bult CJ, Bogue MA. 2014. Mouse phenome database. Nucleic Acids Res. 42:D825–D834.

Hahn MW. 2019. Molecular Population Genetics. New York: Sinauer Associates.

Hernandez Cordero AI, Carbonetto P, Riboni Verri G, Gregory JS, Vandenbergh DJ, *et al.* 2018. Replication and discovery of musculoskeletal QTLs in LG/J and SM/J advanced intercross lines. Physiol Rep. 6:e13561.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, *et al.* 2008. Efficient control of population structure in model organism association mapping. Genetics. 178:1709–1723.

Keele GR, Crouse WL, Kelada SN, Valdar W. 2019. Determinants of QTL mapping power in the realized collaborative cross. G3 (Bethesda). 9:1707–1727.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, *et al.* 2011. FaST linear mixed models for genome-wide association studies. Nat Methods. 8:833–835.

Mackay TF. 2014. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. Nat Rev Genet. 15:22–33.

Ning C, Wang D, Kang H, Mrode R, Zhou L, *et al.* 2018. A rapid epistatic mixed-model association analysis by linear retransformations of genomic estimated values. Bioinformatics. 34: 1817–1825.

Phillips PC. 2008. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet. 9:855–867.

Reifsnyder PC, Churchill G, Leiter EH. 2000. Maternal environment and genotype interact to establish diabesity in mice. Genome Res. 10:1568–1578.

Saul MC, Philip VM, Reinholdt LG, Chesler EJ; Center for Systems Neurogenetics of Addiction. 2019. High-diversity mouse populations for complex traits. Trends Genet. 35:501–514.

Sloan Z, Arends D, Broman KW, Centeno A, Furlotte N, *et al.* 2016. GeneNetwork: framework for web-based genetics. J Open Source Softw. 1:25.

Stich B, Gebhardt C. 2011. Detection of epistatic interactions in association mapping populations: an example from tetraploid potato. Heredity (Edinb). 107:537–547.

Sul JH, Martin LS, Eskin E. 2018. Population structure in genetic studies: confounding factors and mixed models. PLoS Genet. 14: e1007309.

Svenson KL, Gatti DM, Valdar W, Welsh CE, Cheng R, *et al.* 2012. High-resolution genetic mapping using the mouse Diversity Outbred population. Genetics. 190:437–447.

Tyler AL, Donahue LR, Churchill GA, Carter GW. 2016. Weak epistasis generally stabilizes phenotypes in a mouse intercross. PLoS Genet. 12:e1005805.

Tyler AL, Ji B, Gatti DM, Munger SC, Churchill GA, *et al.* 2017. Epistatic networks jointly influence phenotypes related to metabolic disease and gene expression in Diversity Outbred mice. Genetics. 206:621–639.

Tyler AL, Lu W, Hendrick JJ, Philip VM, Carter GW. 2013. CAPE: an R package for Combined Analysis of Pleiotropy and Epistasis. PLoS Comput Biol. 9:e1003270.

Voight BF, Pritchard JK. 2005. Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 1:e32.

Xu Q-S, Liang Y-Z. 2001. Monte Carlo cross validation. Chemom Intell Lab Syst. 56:1–11.

Xu S, Jia Z. 2007. Genomewide analysis of epistatic effects for quantitative traits in barley. Genetics. 175:1955–1963.

Yang H, Bell TA, Churchill GA, De Villena FP-M. 2007. On the subspecific origin of the laboratory mouse. Nat Genet. 39:1100–1107.

Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, *et al.*; the GIANT Consortium. 2011. Genomic inflation factors under polygenic inheritance. Eur J Hum Genet. 19:807–812.

Yao Y, Ochoa A. 2019. Testing the effectiveness of principal components in adjusting for relatedness in genetic association studies. bioRxiv. doi:10.1101/858399.

Communicating editor: T. Mackay