G3
Genes | Genomes | Genetics

Genome-Wide Association Study in Two Cohorts from a Multi-generational Mouse Advanced Intercross Line Highlights the Difficulty of Replication Due to Study-Specific Heterogeneity

Xinzhu Zhou,* Celine L. St. Pierre,[†] Natalia M. Gonzales,[‡] Jennifer Zou,[§] Riyan Cheng,**
Apurva S. Chitre,** Greta Sokoloff,[††] and Abraham A. Palmer**,[‡‡,1]

*Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, CA, 92092, [†]Department of Genetics, Washington University School of Medicine, St. Louis, MO, 63110, [‡]Department of Human Genetics, University of Chicago, IL, 60637, [§]Department of Computer Science, University of California, Los Angeles, CA, 90095, **Department of Psychiatry, [‡‡]Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, 92037 and [††]Department of Psychological & Brain Sciences, University of Iowa, Iowa City, IO, 52242

ORCID IDs: 0000-0002-6432-0033 (X.Z.); 0000-0001-5465-6601 (C.L.St.P.); 0000-0001-5289-0410 (N.M.G.); 0000-0002-2045-3272 (J.Z.); 0000-0003-1709-9214 (A.S.C.); 0000-0003-3651-1980 (G.S.); 0000-0003-3634-0747 (A.A.P.)

ABSTRACT There has been extensive discussion of the "Replication Crisis" in many fields, including genome-wide association studies (**GWAS**). We explored replication in a mouse model using an advanced intercross line (**AIL**), which is a multigenerational intercross between two inbred strains. We re-genotyped a previously published cohort of LG/J x SM/J AIL mice (F_{34}; n = 428) using a denser marker set and genotyped a new cohort of AIL mice (F_{39-43}; n = 600) for the first time. We identified 36 novel genome-wide significant loci in the F_{34} and 25 novel loci in the F_{39-43} cohort. The subset of traits that were measured in both cohorts (locomotor activity, body weight, and coat color) showed high genetic correlations, although the SNP heritabilities were slightly lower in the F_{39-43} cohort. For this subset of traits, we attempted to replicate loci identified in either F_{34} or F_{39-43} in the other cohort. Coat color was robustly replicated; locomotor activity and body weight were only partially replicated, which was inconsistent with our power simulations. We used a random effects model to show that the partial replications could not be explained by Winner's Curse but could be explained by study-specific heterogeneity. Despite this heterogeneity, we performed a mega-analysis by combining F_{34} and F_{39-43} cohorts (n = 1,028), which identified four novel loci associated with locomotor activity and body weight. These results illustrate that even with the high degree of genetic and environmental control possible in our experimental system, replication was hindered by study-specific heterogeneity, which has broad implications for ongoing concerns about reproducibility.

Genome-wide association studies (**GWAS**) in model organisms can use genetically identical cohorts phenotyped under extremely similar conditions, which would be expected to enhance the success of replication.

We sought to investigate replication in model organism GWAS using a mouse advanced intercross line (**AIL**). The use of GWAS in model organisms such as mice (Talbot *et al.* 1999; Demarest *et al.* 2001; Yalcin *et al.* 2004; Valdar *et al.* 2006; Ghazalpour *et al.* 2008; Samocha *et al.* 2010; Churchill *et al.* 2012; Collaborative Cross Consortium 2012; Parker *et al.* 2012, 2016; Svenson *et al.* 2012; Carbonetto *et al.* 2014; Chesler 2014; Coyner *et al.* 2014; Gatti *et al.* 2014; Nicod *et al.* 2016; Hernandez Cordero *et al.* 2018, 2019), rats (Baud *et al.* 2014), chickens (Besnier *et al.* 2011; Johnsson *et al.* 2018), fruit flies (King *et al.* 2012; Mackay *et al.* 2012; Kislukhin *et al.* 2013; Marriage *et al.* 2014; Vonesch *et al.* 2016), *C. elegans* (Doitsidou *et al.* 2016) and various plant species (Rishmawi *et al.* 2017; Cockram and Mackay 2018; Diouf *et al.* 2018) has become increasingly common over the last

decade. These mapping populations can further be categorized as multi-parental crosses, which are created by interbreeding two or more inbred strains, and various outbred populations, in which the founders are of unknown provenance. An F_2 cross between two inbred strains is the prototypical mapping population; however, F_2s provide poor mapping resolution (Parker and Palmer 2011). To improve mapping resolution, Darvasi and Soller (Darvasi and Soller 1995) proposed the creation of advanced intercross lines (**AILs**), which are produced by intercrossing F_2 mice for additional generations. AILs accumulate additional crossovers with every successive generation, leading to a population with shorter linkage disequilibrium (**LD**) blocks, which improves mapping precision, albeit at the expense of power (Parker and Palmer 2011; Gonzales and Palmer 2014).

The longest running mouse AIL was generated by crossing LG/J and SM/J inbred strains, which had been previously selected for large and small body size prior to inbreeding and subsequent intercrossing. We obtained this AIL in 2006 at generation 33 from Dr. James Cheverud (Jmc: LG,SM-G_{33}). Since then, we have collected genotype and phenotype information from multiple generations, including F_{34} (Cheng *et al.* 2010; Lionikas *et al.* 2010; Samocha *et al.* 2010; Parker *et al.* 2011, 2014; Bartnikas *et al.* 2012; Carroll *et al.* 2017; Gonzales *et al.* 2018) and F_{39}-F_{43}. Our previous publications using the F_{34} generation employed a custom Illumina Infinium genotyping microarray to obtain genotypes for 4,593 SNPs (Cheng *et al.* 2010; Parker *et al.* 2014); we refer to this set of SNPs as the 'sparse markers'. Those genotypes were used to identify significant associations for numerous traits, including locomotor activity in response to methamphetamine (Cheng *et al.* 2010), pre-pulse inhibition (Samocha *et al.* 2010), muscle weight (Lionikas *et al.* 2010; Hernandez Cordero *et al.* 2019), body weight (Parker *et al.* 2011), open field (Parker *et al.* 2014), conditioned fear (Parker *et al.* 2014), red blood cell parameters (Bartnikas *et al.* 2012), and muscle weights (Carroll *et al.* 2017). Although not previously published, we also collected phenotype information from the F_{39}-F_{43} generations, including body weight, fear conditioning, locomotor activity in response to methamphetamine, and the light dark test for anxiety.

While the prior GWAS using the F_{34} generation detected many significant loci, the sparsity of the markers likely precluded the discovery of some true loci and also made it difficult to clearly define the boundaries of the loci that we did identify. For example, Parker *et al.* conducted an integrated analysis of F_2 and F_{34} AILs (Parker *et al.* 2011). One of their body weight loci spanned from 87.93–102.70 Mb on chromosome 14. Denser markers might have more clearly defined the implicated region.

In the present study, we used genotyping-by-sequencing (**GBS**), which is a reduced-representation sequencing method (Davey *et al.* 2011; Elshire *et al.* 2011; Fitzpatrick *et al.* 2013), to obtain a much denser set of SNPs in the F_{34} cohort and, for the first time, genotyped mice from the F_{39}-F_{43} generations. With this denser set of SNPs, we attempted to identify novel loci in the F_{34} cohort that were not detected using the sparse SNPs. We also performed GWAS using the mice from the F_{39}-F_{43} AILs. We explored whether imputation from the array SNPs could have provided the additional coverage we obtained using the denser GBS genotypes. Because F_{39}-F_{43} AILs are the direct descendants of the F_{34}, they are uniquely suited to serve as a replication population for GWAS in the F_{34} generation. For the subset of traits measured in both cohorts, we attempted to replicate the results discovered in one cohort in the other. To set our expectations for replication, we performed simulations to estimate the power for these replication studies. Because the actual rate of replication was lower than predicted by the

power analysis, we used a random effects model to evaluate the role of Winner's Curse and study-specific heterogeneity in the low rate of replication. Finally, we also performed a mega-analysis on a subset of traits common to both cohorts.

MATERIALS AND METHODS

Animals
All mice used in this study were members of the LG/J x SM/J AIL that was originally created by Dr. James Cheverud (Loyola University Chicago, Chicago, IL). This AIL has been maintained in the Palmer laboratory since generation F_{33}. Age and exact number of animals tested in each phenotype are described in Table S1. Several previous publications (Samocha *et al.* 2010; Cheng *et al.* 2010; Parker *et al.* 2014; Lionikas *et al.* 2010; Carroll *et al.* 2017; Parker *et al.* 2011; Bartnikas *et al.* 2012) have reported association analyses of the F_{34} mice (n = 428). No prior publications have described the F_{39}-43 generations (n = 600). The sample size of F_{34} mice reported in this study (n = 428) is smaller than that in previous publications of F_{34} (n = 688) because we only genotyped a subset of F_{34} animals using GBS.

F_{34}, F_{39}-43 Phenotypes
All phenotypes are listed in Table S1. We have previously described the phenotyping of F_{34} animals for locomotor activity and locomotor response to methamphetamine (Cheng *et al.* 2010), fear conditioning (Parker *et al.* 2014), open field (Parker *et al.* 2014), coat color, body weight (Parker *et al.* 2011), complete blood counts (Bartnikas *et al.* 2012), heart and tibia measurements (Lionikas *et al.* 2010), muscle weight (Lionikas *et al.* 2010). Iron content in liver and spleen, which have not been previously reported in these mice, was measured by atomic absorption spectrophotometry, as described in Gardenghi *et al.* (Gardenghi *et al.* 2007) and Graziano, Grady and Cerami (Graziano *et al.* 1974). Although the phenotyping of F_{39}-43 animals has not been previously reported, we followed previously published protocols for locomotor activity and locomotor response to methamphetamine (Cheng *et al.* 2010), coat color, body weight (Parker *et al.* 2011), and light/dark test for anxiety (Sittig *et al.* 2016). We point out here that even though "locomotor activity" was measured in both the F_{34} and F_{39}-43 using the Versamax software (AccuScan Instruments, Columbus, OH), "open field" in the F_{34} cohort was also measured using Versamax, whereas "open field" in the F_{39}-43 cohort was measured using the EthoVision XT software (Noldus system; (Noldus *et al.* 2001)). Because there are meaningful differences in these experimental procedures, we did not attempt to use "open field" data for replication. In summary, we performed GWAS on all traits collected in individual cohorts. For the replication analysis between the F_{34} and F_{39}-43 cohorts, we only directly compared a number of traits that had been measured in both cohorts: body weight, two Mendelian coat color traits (albino and agouti), and three locomotor activity traits (locomotor activity on day 1 and on day 2, and activity on day 3 following a methamphetamine injection).

F_{34} AIL Array Genotypes
F_{34} animals had been genotyped on a custom SNP array on the Illumina Infinium platform (Cheng *et al.* 2010; Parker *et al.* 2014), which yielded a set of 4,593 SNPs on autosomes and X chromosome that we refer to as 'sparse SNPs'.

F_{34} and F_{39}-43 GBS Genotypes
F_{34} and F_{39}-43 animals were genotyped using genotyping-by-sequencing (**GBS**), which is a reduced-representation genome sequencing

method (Parker *et al.* 2016; Gonzales *et al.* 2017). We used the same protocol for GBS library preparation that was described in Gonzales *et al.* (Gonzales *et al.* 2017). We called GBS genotype probabilities using ANGSD (Korneliussen *et al.* 2014). GBS identified 1,667,920 autosomal and 43,015 X-chromosome SNPs. To fill in missing genotypes at SNPs where some but not all mice had calls, we ran within-sample imputation using Beagle v4.1, which generated hard call genotypes as well as genotype probabilities (Browning and Browning 2007). After imputation, only SNPs that had dosage $r^2 > 0.9$ were retained. We removed SNPs with minor allele frequency < 0.1 and SNPs with $P < 1.0 \times 10^{-6}$ in the Chi-square test of Hardy–Weinberg Equilibrium (**HWE**) (Table S2). All phenotype and GBS genotype data are deposited in GeneNetwork2 (http://gn2.genenetwork.org/).

QC of individuals
We have found that large genetic studies are often hampered by cross-contamination between samples and sample mix-ups. We used four features of the data to identify problematic samples: heterozygosity distribution, proportion of reads aligned to sex chromosomes, pedigree/kinship, and coat color. We first examined heterozygosity across autosomes and removed animals where the proportion of heterozygosity was more than 3 standard deviations from the mean (Figure S1). Next, we sought to identify animals in which the recorded sex did not agree with the sequencing data. We compared the ratio of reads mapped to the X and Y chromosomes. The 95% CI for this ratio was 196.84 to 214.3 in females and 2.13 to 2.18 in males. Twenty-two F_{34} and F_{39-43} animals were removed because their sex (as determined by reads ratio) did not agree with their recorded sex; we assumed this discrepancy was due to sample mix-ups. To further identify mislabeled samples, we calculated kinship coefficients based on the full AIL pedigree using QTLRel. We then calculated a genetic relatedness matrix (**GRM**) using IBDLD (Abney 2008; L. Han and Abney 2011), which estimates identity by descent using genotype data. The comparison between pedigree kinship relatedness and genetic kinship relatedness identified seven pairs of animals that showed obvious disagreement between kinship coefficients and the GRM, these animals were excluded from further analysis. Lastly, we excluded 14 F_{39-43} animals that showed discordance between their recorded coat color and their genotypes at markers flanking *Tyr*, which causes albinism in mice. The numbers of animals filtered at each step are listed in Table S2. Some animals were detected by more than one QC step, substantiating our evidence that these samples were erroneous.

At the end of SNP and sample filtering, we had 59,561 autosomal and 831 X chromosome SNPs in F_{34}, 58,966 autosomal and 824 X chromosome SNPs in F_{39-43}, and 57,635 autosomal and 826 X chromosome SNPs in the combined F_{34} and F_{39-43} set (Table S2). GBS genotype quality was estimated by examining concordance between the 66 SNPs that were present in both the array and GBS genotyping results (Figure S3).

LD decay
Average LD (r^2) was calculated using allele frequency matched SNPs (MAF difference < 0.05) within 100,000 bp distance, as described in Parker *et al.* (Parker *et al.* 2016).

Imputation to LG/J and SM/J reference panels
F_{34} array genotypes (n = 428) and F_{34} GBS genotypes (n = 428) were imputed to LG/J and SM/J whole genome sequence data (Nikolskiy *et al.* 2015) using BEAGLE (Browning and Browning 2007). For F_{34} array imputation, we used a large window size (100,000 SNPs and

45,000 SNPs overlap). Imputation to reference panels yielded 4.3 million SNPs for F_{34} array and F_{34} GBS imputed sets. Imputed SNPs with $r^2 > 0.9$, MAF > 0.1, HWE p-value $> 1.0 \times 10^{-6}$ were retained, resulting in 4.1M imputed F_{34} GBS SNPs and 4.3M imputed F_{34} array SNPs.

Genome-wide association analysis (GWAS)
We used the linear mixed model, as implemented in GEMMA (Zhou and Stephens 2012), to perform a GWAS that accounted for the complex familial relationships among the AIL mice (Cheng *et al.* 2010; Gonzales *et al.* 2017). We used the leave-one-chromosome-out (**LOCO**) approach to calculate the GRM, which effectively circumvented the problem of proximal contamination (Cheng *et al.* 2013). We used the univariate linear mixed model described in Zhou and Stephens (Zhou and Stephens 2012):

$$y = W\alpha + x\beta + u + \varepsilon; u \sim MVN_n\left(0, \lambda\tau^{-1}K\right), \varepsilon \\ \sim MVN_n\left(0, \tau^{-1}I_n\right),$$

where y is a n-vector of traits for n individuals; W is a n×c matrix of covariates (fixed effects); α is a c-vector of the corresponding coefficients; x is an n-vector of genotypes; β is the effect size of the genotype; u is an n-vector of random effects; ε is an n-vector of errors; τ^{-1} is the variance of the residual errors; λ is the ratio between the two variance components; K is a known n × n relatedness matrix and I_n is an n × n identity matrix. MVN_n stands for the n-dimensional multivariate normal distribution (Zhou and Stephens 2012).

Separate GWAS were performed using the F_{34} array genotypes, the F_{34} GBS genotypes, and the F_{39-43} GBS genotypes. Apart from coat color (binary trait), raw phenotypes were quantile normalized prior to analysis. Coat color traits were coded as follows: albino: 1 = white, 0 = non-white; agouti: 1 = tan, 0 = black, NA = white. Because F_{34} AIL had already been studied, we used the same covariates as described in Cheng *et al.* (Cheng *et al.* 2010) in order to examine whether our array and GBS GWAS would replicate their findings. We included sex and body weight as covariates for locomotor activity traits (see covariates used in (Cheng *et al.* 2010)) and sex, age, and coat color as covariates for fear conditioning and open field test in F_{34} AILs (see covariates used in (Parker *et al.* 2014)). We used sex and age as covariates for all other phenotypes. Covariates for each analysis are shown in Table S1. Finally, we performed mega-analysis of F_{34} and F_{39-43} animals (n = 1,028) for body weight, coat color, and locomotor activity, since these traits were measured in the same way in both cohorts. We quantile transformed all continuous phenotypes in each cohort and then combined the transformed phenotypes for the mega-analysis (coat color traits were not quantile normalized because they are binary).

Identifying dubious SNPs
Some significant SNPs in the F_{34} GWAS were dubious because the flanking SNPs, which would have been expected to be in high LD with the significant SNP (a very strong assumption in an AIL), did not have high -log10(p) values. We only examined SNPs that obtained significant p-values; close examinations revealed that these SNPs had dubious ratios of heterozygotes to homozygotes calls and had corresponding HWE p-values that were close to our 1.0×10^{-6} threshold (Table S3). We chose the 1.0×10^{-6} as the filter threshold of the HWE p-values based on a gene-dropping exercise. We used the F_{33-34} family pedigree and the F_{34} genetic map to simulate the genotypes in F_{34} (QTLRel; (Cheng *et al.* 2011)). The p-value of the chi-square test for Hardy-Weinberg equilibrium in the simulated F_{34} population was

7.24329×10^{-06}, which was close to the HWE threshold used in Gonzales *et al.* (Gonzales *et al.* 2018). To avoid counting these as novel loci, we removed those SNPs prior to summarizing our results as they likely reflected genotyping errors.

Selecting independent significant SNPs

To identify independent "lead loci" among significant GWAS SNPs that surpassed the significance threshold, we used the LD-based clumping method in PLINK v1.9. We empirically chose clumping parameters ($r^2 = 0.1$ and sliding window size = 12,150kb) that gave us a conservative set of independent SNPs (Table S4). For the coat color phenotypes, we found that multiple SNPs remained significant even after LD-based clumping, presumably due to the extremely significant associations at these Mendelian loci. In these cases, we used a stepwise model selection procedure in GCTA (Yang *et al.* 2011) and performed association analyses conditioning on the most significant SNPs.

Significance thresholds

We used MultiTrans to set significance thresholds for GWAS (B. Han *et al.* 2009; Joo *et al.* 2016). MultiTrans is a method that assumes multivariate normal distribution of the phenotypes, which in LMM models, contain a covariance structure due to various degrees of relatedness among individuals. We were curious to see whether MultiTrans produced significance thresholds that were different from the thresholds we obtained from a standard permutation test ('naïve permutation' as per Cheng *et al.* (Cheng *et al.* 2013)). We performed 1,000 permutations using the F_{34} GBS genotypes and the phenotypic data from locomotor activity (days 1, 2, and 3). We found that the 95th percentile values for these permutations were 4.65, 4.79, and 4.85, respectively, which were very similar to 4.85, the threshold obtained from MultiTrans using the same data. Thus, the thresholds presented here were obtained from MultiTrans but are similar (if anything slightly more conservative) to the thresholds we would have obtained had we used permutation. Because the effective number of tests depends on the number of SNPs and the specific animals used in GWAS, we obtained a unique adjusted significance threshold for each SNP set in each animal cohort (Table S5).

Credible set analysis

We followed the method described in (Wellcome Trust Case Control Consortium *et al.* 2012). Credible set analysis is a Bayesian method of selecting an interval of SNPs that are likely to contain the causal SNPs; we used LD r^2 threshold = 0.8, posterior probability =0.99. The R script could be found on GitHub: https://github.com/hailianghuang/FM-summary/blob/master/getCredible.r

Power analysis

To estimate the power of replication of a SNP from the discovery set in the replication set, we simulated GWAS with 50 varying effect sizes for the discovery SNP using the LMM model. We first fit the trait in a null model (*i.e.*, no genotype effect), and obtained estimates of model parameters including the intercept and the genetic variance component. Using these model parameters, we added the genotype effect to the random numbers generated from the null model to recreate a trait. For each simulated effect size, we scanned every simulated trait 2,500 times and examined the ratio of association tests whose test statistics surpassed the significance thresholds (both the genome-wide significance threshold for the cohort and the nominal p-value of 0.05).

Replication analysis between F_{34} and F_{39-43} GWAS studies

We modeled the replication between F_{34} and F_{39-43} GWAS studies using two random effects models (Zou *et al.* 2019). Both models take as input a set of z-scores for variants computed from an association study ("summary statistics").

The **WC** model accounts only for Winner's Curse. We assume that there is a shared genetic effect (λ) that is responsible for the observed association signal in both studies. To model random noise contributing to Winner's Curse, we model the summary statistics for each variant k from the discovery and replication studies as normally distributed random variables ($s_k^{(1)} \sim N(\lambda, \ 1)$ and $s_k^{(2)} \sim N(\lambda, \ 1)$, respectively). We define the prior probability of the true genetic effect to be $\lambda \sim N(0, \ \sigma_g^2)$, where the variance in the true genetic effect is learned through a maximum likelihood procedure. We correct for the effect of Winner's Curse in the discovery study by computing the conditional distribution of the replication summary statistic given the discovery summary statistic.

The **WC+C** model accounts for Winner's Curse and study-specific heterogeneity. In this model, we partition the total effect sizes observed into genetic effects (λ) and study-specific effects ($\delta^{(1)}$ and $\delta^{(2)}$). We model the statistics for each variant k from the initial and discovery studies as normally distributed random variables ($s_k^{(1)} \sim N(\lambda + \delta^{(1)}, \ 1)$ and $s_k^{(2)} \sim N(\lambda + \delta^{(2)}, \ 1)$, respectively). In addition to the prior on the genetic effect defined in the WC model, we define the prior probabilities of the study-specific effects to be $\delta^{(1)} \sim N(0, \ \sigma_{c_1}^2)$, and $\delta^{(2)} \sim N(0, \ \sigma_{c_2}^2)$, where the variance parameters are learned through a maximum likelihood procedure. We correct for the effect of Winner's Curse in the discovery study and study-specific effects by computing the conditional distribution of the replication summary statistic given the discovery summary statistic.

We applied each of these models once using F_{34} as the discovery study and once using F_{39-43} as the discovery study. We used the genome-wide significance thresholds in Table S5 to identify variants in each discovery study and used the results as input to the random effects models. We then used a Bonferroni corrected threshold ($P = 0.05/M$) for the replication study, where M is the number of genome-wide significant variants in the initial study. We computed the "empirical replication rate" as the proportion of variants passing the genome-wide significant threshold in the discovery study that also passed this Bonferroni corrected threshold in the replication study. Since the estimation of the model parameters requires at least two variants, we only applied this method to phenotypes with at least two genome-wide significant variants in the discovery study.

To assess how well the **WC** and **WC+C** models explained the observed patterns of replication, we computed the predicted replication rates under each model. For each variant that passed the genome-wide significant threshold in the discovery study, we used the conditional distributions previously learned to compute the probability that the variant passed the Bonferroni corrected threshold in the replication study. For each phenotype, we computed the average of these predicted replication rates and compared this average to the empirical replication rates.

Genetic correlation and heritability estimates between F_{34} and F_{39-43} phenotypes

Locomotor activity, body weight, and coat color traits had been measured in both F_{34} and F_{39-43} populations. We calculated both SNP heritability and genetic correlations between F_{34} and F_{39-43} animals

using GCTA-GREML analysis and GCTA bivariate GREML analysis (Yang *et al.* 2011).

LocusZoom Plots

LocusZoom plots were generated using the standalone implementation of LocusZoom (Pruim *et al.* 2010), using LD scores calculated from PLINK v.1.9–ld option and mm10 gene annotation file downloaded from UCSC Genome Browser.

Data availability

All relevant data are within the paper and its Supporting Information files. Genotypes and phenotypes of F_{34} ("LGSM AI G34 Palmer (Array)": GN655; "LGSM AI G34 Palmer (GBS)": GN656), F_{39-43} ("LGSM AI G39-43 Palmer (GBS)": GN657), and mega-analysis cohort ("LGSM AI G34 G39-43 Palmer (GBS)": GN654) of AIL are uploaded to GeneNetwork2 (http://gn2.genenetwork.org/). Code used to perform the analyses is included in the supplementary materials on figshare: https://doi.org/10.25387/g3.11674221.

RESULTS

We used 214 males and 214 females from generation F_{34} (Aap:LG,SM-G34) and 305 males and 295 females from generations F_{39-43}. For the F_{34} AIL 79 traits were available from previous published and unpublished work; for the F_{39-43} AIL 49 unpublished traits were available (Table S1). F_{34} mice had been previously genotyped on a custom SNP array (Cheng *et al.* 2010; Parker *et al.* 2014). The average minor allele frequency (**MAF**) of those 4,593 array SNPs was 0.388 (Figure 1). To obtain a denser set of SNP markers, we used GBS in F_{34} and F_{39-43} AIL mice. Since data on the F_{39-43} AIL mice had been collected over the span of approximately two years, we carefully considered the possibility of sample contamination and sample mislabeling (Toker *et al.* 2016) and removed these samples (see Methods; Figure S1 and S2). The final SNP sets included 60,392 GBS-derived SNPs in 428 F_{34} AIL mice, 59,790 GBS-derived SNPs in 600 F_{39-43} AIL mice, and 58,461 GBS-derived SNPs that existed in both F_{34} and F_{39-43} AIL mice (Table S2). The MAF for the GBS SNPs was 0.382 in F_{34}, 0.358 in F_{39-43}, and 0.370 in F_{34} and F_{39-43} (Figure 1). There were 66 SNPs called from our GBS data that were also present on the genotyping array. The genotype concordance rate for those 66 SNPs, which reflects the sum of errors from both sets of genotypes, was 95.4% (Figure S3). We found that LD decay rates using F_{34} array, F_{34} GBS, F_{39-43} GBS, and F_{34} and F_{39-43} GBS genotypes were generally similar to one another, though levels of LD using the GBS genotypes appear to be slightly reduced in the later generations of AILs (Figure S4).

GBS genotypes produced more significant associations than array genotypes in F_{34}

We used a linear mixed model (**LMM**) as implemented in GEMMA (Zhou and Stephens 2012) to perform GWAS. We used the leave-one-chromosome-out (**LOCO**) approach to address the problem of proximal contamination, as previously described (Listgarten *et al.* 2012; Cheng *et al.* 2013; Yang *et al.* 2014; Gonzales *et al.* 2017). We performed GWAS using both the sparse array SNPs and the dense GBS SNPs to determine whether additional SNPs would produce more genome-wide significant associations. Autosomal and X chromosome SNPs were included in all GWAS. We obtained a significance threshold for each SNP set using MultiTrans (B. Han *et al.* 2009; Joo *et al.* 2016).To select independently associated loci ("lead loci"), we used an LD-based clumping method implemented in PLINK to group SNPs that passed the adjusted genome-wide significance thresholds over a large genomic region flanking the index SNP (Purcell *et al.* 2007). Applying the most

stringent clumping parameters ($r^2 = 0.1$ and sliding window size = 12,150kb, Table S4), we identified 109 significant lead loci in 49 out of 79 F_{34} phenotypes using the GBS SNPs (Table S7). In contrast, we identified 83 significant lead loci in 45 out of 79 F_{34} phenotypes using the sparse array SNPs (Table S6, Table S7). Among the loci identified in the F_{34}, 36 were uniquely identified using the GBS genotypes, whereas 11 were uniquely identified using the array genotypes. These unique loci could be explained by the disparity of the marker density between the GBS and array genotypes. Some unique loci captured haplotype blocks that were not picked up in the other SNP set. Other unique loci were only slightly above the significance threshold in one SNP set where the corresponding loci in the other SNP set had sub-threshold significance (*i.e.*, p-value $\sim 10^{-5}$ but below the significance threshold of the cohort; Table S7). Overall, GBS SNPs consistently yielded more significant lead loci compared to array SNPs regardless of the clumping parameter values (Table S4), indicating that a dense marker panel was able to detect more association signals compared to a sparse marker panel.

To determine the boundaries of each locus, we performed a Bayesian-framework credible set analysis, which estimated a posterior probability for association at each SNP (r^2 threshold = 0.8, posterior probability threshold = 0.99; (Wellcome Trust Case Control Consortium *et al.* 2012)). The physical positions of the SNPs in the credible set were used to determine the boundaries of each locus. As expected, the greater density of the GBS genotypes allowed us to better define each interval. For instance, the lead locus at chr17:27130383 was associated with distance traveled in periphery in the open field test in F_{34} AILs (Figure 2). However, no SNPs were genotyped between 26.7 and 28.7 Mb in the array SNPs, which makes the size of this LD block ambiguous. In contrast, the Locus-Zoom plot portraying GBS SNPs in the same region shows that SNPs in high LD with chr17:27130383 are between 27 Mb and 28.3 Mb. The more accurate definition of the implicated intervals allowed us to better refine the list of the coding genes and non-coding variants associated with the phenotype (Table S6).

In our prior studies using the sparse marker set, we did not attempt to increase the number of available markers by using imputation. Therefore, we examined whether the disparity between the numbers of loci identified by the two SNP sets could be resolved by imputation, which should increase the number of markers available for GWAS. We used LG/J and SM/J whole genome sequencing data as reference panels (Nikolskiy *et al.* 2015) and performed imputation on array and GBS SNPs using Beagle v4.1 (Browning and Browning 2007). After QC filtering, we obtained 4.3M SNPs imputed from the array SNPs and 4.1M SNPs imputed from the GBS SNPs. More imputed GBS SNPs were filtered out because GBS SNPs were called from genotype probabilities, thus introducing uncertainty in imputed SNPs. We found that imputed array genotypes and imputed GBS genotypes did not meaningfully increase the number of loci discovered, presumably because the utility of imputation is inherently limited in a two-strain cross.

Under a polygenic model where a large number of additive common variants contribute to a complex trait, heritability estimates could be higher when more SNPs are considered (Yang *et al.* 2017). Given that there were more GBS SNPs than array SNPs, we used autosomal SNPs to examine whether GBS SNPs would generate higher SNP heritability estimates compared to the sparse array SNPs. Heritability estimates were similar for the two SNP sets, with the exception of agouti coat color, which showed marginally greater heritability for the GBS SNPs (Figure S5; Table S8). Our results show that while the denser GBS SNP set was able to identify more
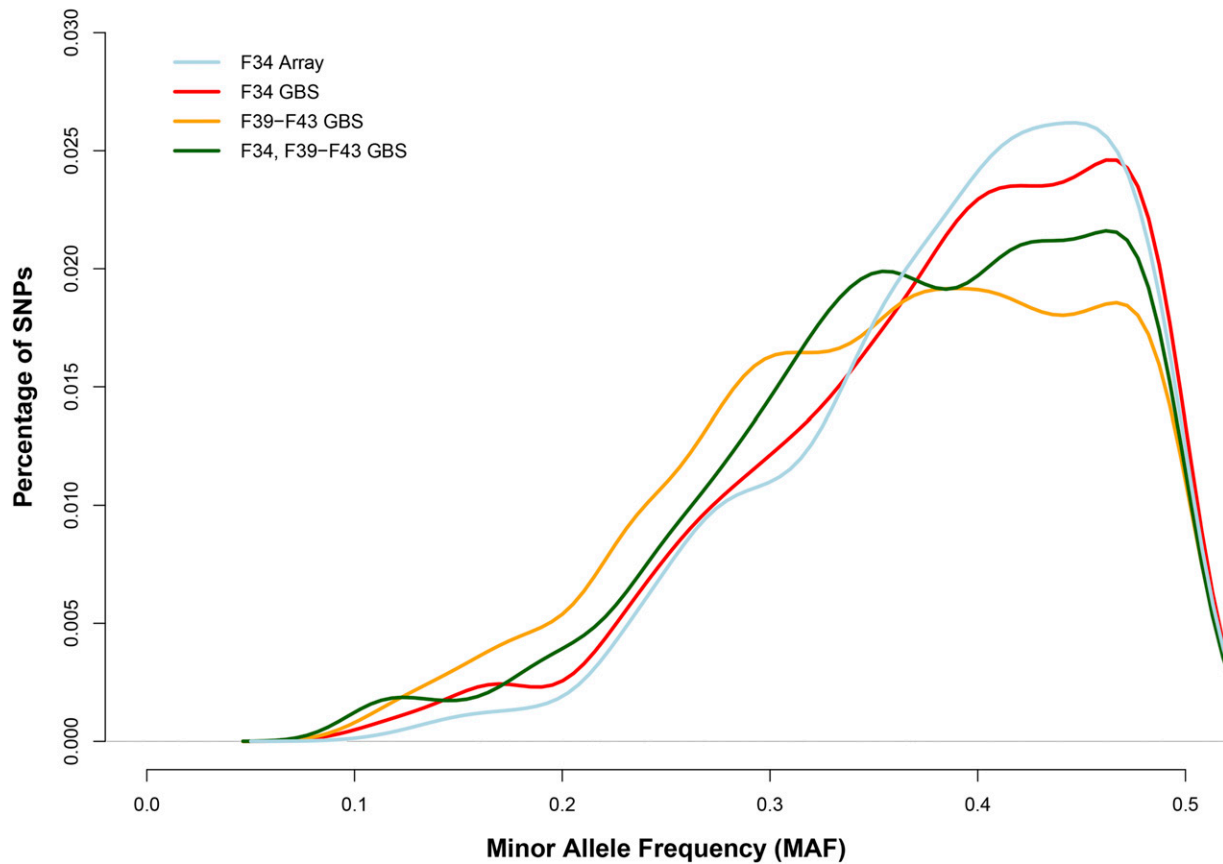
Figure 1 Minor allele frequency (MAF) distributions for F_{34} array, F_{34} GBS, F_{39}-F_{43} GBS, and F_{34} and F_{39}-F_{43} GBS SNP sets. The average MAF of those 4,593 array SNPs was 0.388; the average MAF of the 60,392 GBS-derived SNPs in 428 F_{34} AIL mice was 0.382; the average MAF of the 59,790 GBS-derived SNPs in 600 F_{39-43} AIL mice was 0.358; the average MAF of the 58,461 GBS-derived SNPs that existed in both F_{34} and F_{39-43} AIL mice was 0.370 (Table S2). MAF distributions are highly comparable between AIL generations.

genome-wide significant loci, greater SNP density did not improve the polygenic signal.

Partial replication of loci identified in F_{34} or F_{39-43} and mega-analysis

We identified 25 genome-wide significant loci for 21 phenotypes in the F_{39-43} cohort (Table S9). A subset of those traits: coat color, body weight, and locomotor activity, were also phenotyped in the F_{34} AILs. To assess replication, we determined whether the loci that were significant in one cohort (either F_{34} or F_{39-43}) would also be significant in the other. We termed the cohort in which a locus was initially discovered as its "discovery set" and the cohort we attempted replication in as the "replication set" (Table 1). Coat color phenotypes (both albino and agouti) are Mendelian traits and thus served as positive controls. All coat color and body weight loci were replicated. The three body weight loci identified in the F_{34} were replicated at nominal levels of significance ($P < 0.05$) in F_{39-43}; similarly, one body weight locus identified in F_{39-43} was replicated in F_{34} ($P < 0.05$). However, none of the locomotor activity loci were replicated in the reciprocal (replication) cohorts.

We found that using a broader definition of an association region rather than a single SNP did not improve replication between the F_{34} cohort and the F_{39-43} cohorts. Confidence intervals (*e.g.*, (Baud *et al.* 2013; Nicod *et al.* 2016)) and the LOD support interval (Conneally *et al.* 1985; Lander and Botstein 1989) have been used to define a QTL. LOD

support interval is very sensitive to the density of the SNPs where sparse markers would produce misleadingly large support intervals. In contrast, the credible set interval is an estimate of the posterior probability for association at markers neighboring the discovery SNP, and thus defines the size of the association region. As a result, we extended the replication comparison from the discovery SNP position to the credible set interval. We found that in the replication cohort, the p-value at the discovery SNP and that at the top SNP within the credible set interval (defined by the discovery QTL) were generally similar (Table S10). The replication of the locus chr14.79312393 (discovered in the F_{34} cohort) in the F_{39-43} cohort was more successful using the discovery QTL region defined by the credible set interval; the p-value at the top SNP within the credible set interval was noticeably more significant (chr14.82586326; p-value = 1.48×10^{-6}) than the p-value at the discovery SNP (chr14.79312393; p-value = 0.0237; Table S10). Our results suggest that for the most part, the discovery SNP accurately represented the association strength of the loci, presumably because of its strong linkage with the neighboring SNPs. In our case, defining a QTL region by the credible set interval did not increase the count of replicated sites between the two cohorts.

We then considered the more liberal "sign test", a statistical method to test for consistent differences between pairs of observations, to determine whether the directions of the effect (beta) of the coat color, body weight and activity loci were in the same direction between the
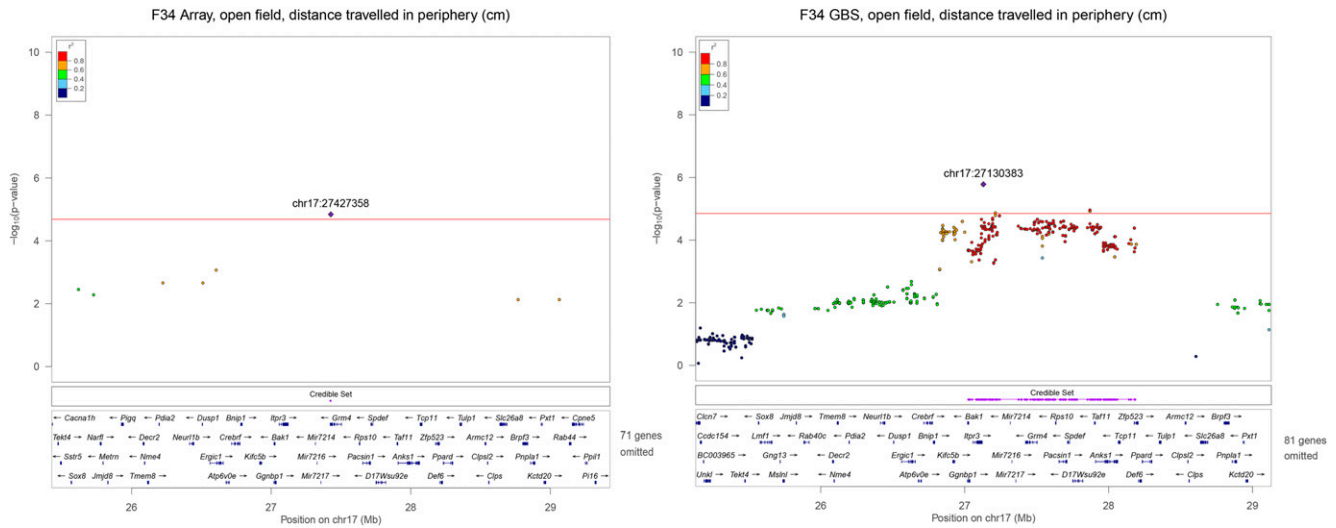
Figure 2 Significant loci on chromosome 17 for open field, distance traveled in periphery in F_{34} AIL. As exemplified in this pair of LocusZoom plots, GBS SNPs defined the boundaries of the loci much more precisely than array SNPs. GBS SNPs that are in high LD ($r^2 > 0.8$, red dots) with lead SNP chr17:27130383 resides between 27 ∼28.3 Mb. In contrast, too few SNPs are present in the array plot to draw any definitive conclusion about the boundaries or LD pattern in this region. Purple track shows the credible set interval. LocusZoom plots for all loci identified in this paper are in Figure S8.

discovery and replication cohorts. Specifically, we compared whether the sign (direction) of the beta estimates are consistently above or below zero. We found that 11 of 12 comparisons passed this much less stringent test of replication. The one locus (at chr15.67627183) that did not pass the sign test was the locomotor locus "discovered" in F_{39-43} (Table 1).

In light of the failure to replicate the locomotor activity findings, we conducted a series of 2,500 simulations per trait to estimate the expected power of our replication cohorts. For each phenotype we used the kinship relatedness matrix and variance components estimated from the replication set. For the coat color traits, we found that we had 100% power to replicate the association at either genome-wide significant levels or the more liberal $p < 0.05$ threshold (Figure S6). For body weight and locomotor activity, power to replicate at a genome-wide significance threshold ranged from 20 to 85%, whereas power to replicate at the $p < 0.05$ threshold was between 80% and 100% (Figure S6). These power estimates were inconsistent with our empirical observations for the locomotor activity traits, none of which replicated at even the $p < 0.05$ threshold, where we should have had almost 100% power (Table 1; Figure S6). However, our power simulations did not account for Winner's Curse (Zöllner and Pritchard 2007) or study-specific heterogeneity (Zou *et al.* 2019).

To determine whether these factors could explain the lower than expected rate of replication, we applied a statistical framework that jointly models Winner's Curse and study-specific heterogeneity in two GWAS studies of the same phenotype (Zou *et al.* 2019). This framework proposes two random effects models. The first model (**WC**) only accounts for Winner's Curse, while the second model accounts for both Winner's Curse and study-specific heterogeneity due to confounding (**WC+C**). In this context, we define confounding as any biological or technical effect present in one study but not the other. We applied each of these models once using F_{34} as the discovery study and once using F_{39-43} as the discovery study. The models can be used to assess how well Winner's Curse explains the observed levels of replication. For example, when F_{34} is used as the replication study for the albino coat color phenotype, the expected value of the replication summary statistics

after accounting for winner's curse is the same as the expected value after accounting for Winner's Curse and confounding (Figure S7). While the 95% confidence intervals for the **WC+C** model are larger than the **WC** model, both models seem to explain the observed data well. However, when F_{34} is used as the discovery study for the locomotor activity on day 1 or body weight, the **WC+C** model explains the data better than the **WC** model.

In order to quantitatively assess how well each of these models explain the observed patterns of replication, we computed the predicted replication rates under each model (Methods) and compared these with the empirical replication rates. In this analysis, we defined the empirical replication rate to be the proportion of variants passing the genome-wide significance threshold in the discovery study that also pass the Bonferroni corrected threshold in the replication study. We used this definition of replication for this analysis instead of replication of lead SNPs to allow for a larger number of variants to be included in the model fitting process. For all phenotypes tested, the **WC** model predicts that all the variants passing the genome-wide significance threshold in the discovery study should pass the Bonferroni corrected threshold in the replication study, which is dramatically different from the observed replication of body weight and locomotor activity on day 1 and 2 phenotypes (Table 2). While the replication in the agouti coat color phenotype is not well predicted by the **WC+C** model, this may be due to the fact that the agouti phenotype is a dominant trait, while our model assumes additive allele effects. These results suggest that the sample sizes are sufficiently large that Winner's Curse cannot account for the lack of replication. However, in these cases, the **WC+C** model has predicted replication rates that are much closer to the true (observed) values, indicating that the lack of replication in these phenotypes is more likely to be due to study-specific heterogeneity that is potentially caused by confounding.

We evaluated whether or not the traits showed genetic correlations across the two cohorts. High genetic correlations would indicate a high degree of additive genetic effect that is shared between the two cohorts, and the low genetic correlations would indicate limited potential for replication. We used all autosomal SNPs to calculate genetic correlations between the F_{34} and F_{39-43} generations for body weight, coat color, and
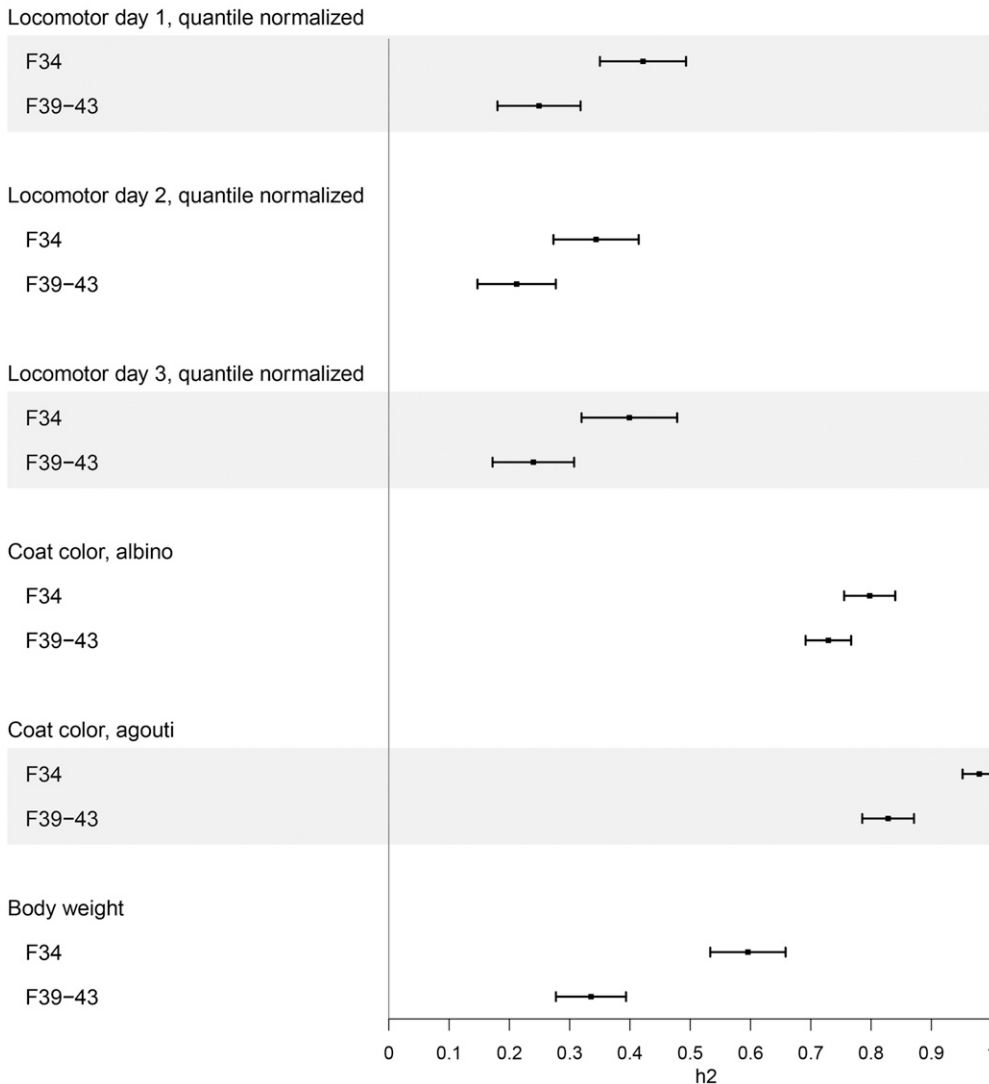
Phenotype

Figure 3 SNP-heritability estimates in F_{34} and F_{39-43} AILs. Square dots represent the SNP heritability estimated by the GCTA-GREML analysis (Yang *et al.* 2011). The whiskers flanking the square dots show the $\pm 1 \times$ standard error of the heritability estimate. All heritability estimates are highly significant ($p < 1.0 \times 10^{-05}$; see Table S12).

locomotor activity phenotypes (Table S11), using GCTA-GREML (Yang *et al.* 2011). Albino and agouti coat color, body weight and locomotor activity on days 1 and 2 were highly genetically correlated ($r_G > 0.7$; Table S11). In contrast, locomotor activity on day 3 showed a significant but weaker genetic correlation ($r_G = 0.577$), perhaps reflecting variability in the quality of the methamphetamine injection, which were only given on day 3. Overall, these results suggest that genetic influences on these traits were largely similar in the two cohorts; however, the genetic correlations were less than 1, suggesting an additional barrier to replication that was not accounted for in our power simulations.

We also calculated the SNP heritability for all traits using GCTA. SNP heritability was consistently lower in the F_{39-43} cohort compared to the F_{34} cohort, including the Mendelian traits of coat color. The $\pm 1 \times$ standard error intervals of the F_{34} and F_{39-43} SNP heritability estimates for the coat color trait albino overlapped. This observation indicates that SNP heritability for albino in the two cohorts is comparable. In contrast, the $\pm 1 \times$ standard error intervals of the F_{34} and F_{39-43} SNP

heritability estimates for the coat color trait agouti did not overlap. We could not explain the differential SNP heritability for the binary trait agouti in the two cohorts. The lower SNP heritability in F_{39-43} for the rest of the quantitative traits could be a result of increased experimental variance (Figure 3; Table S12; (Falconer 1960; Lynch and Walsh 1996; Mhyre *et al.* 2005; Zöllner and Pritchard 2007; Visscher *et al.* 2008; Zaitlen and Kraft 2012)).

Due to the relatively high genetic correlations (Table S11), we suspected that a mega-analysis using the combined sample set would allow for the identification of additional loci; indeed, mega-analysis identified four novel genome-wide significant associations (Figure 4; Table S13). The significance level of five out of six loci identified by the mega-analysis was greater than that in either individual cohort. For instance, the p-values obtained by mega-analysis for chr14:82672838 (p-value = 7.93×10^{-9}) for body weight were lower than the corresponding p-values for the same loci for F_{34} (chr14:79312393, p-value = 7.53×10^{-6}) and F_{39-43} (chr14.82586326, p-value = 2.63×10^{-6}; Table S13; Table 1).
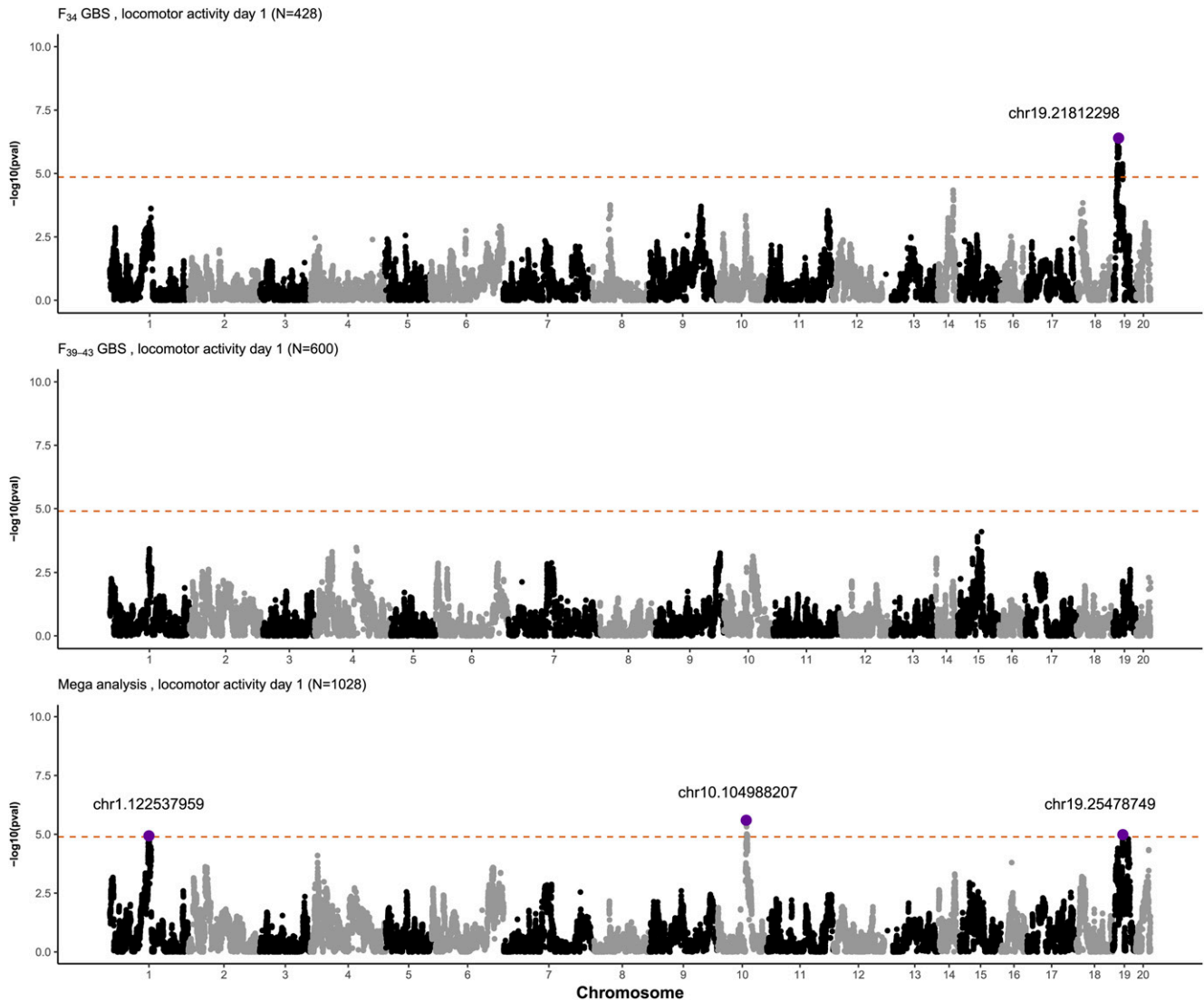
Figure 4 Manhattan plots comparing F_{34} GBS, F_{39-43} GBS, and mega-analysis on locomotor day 1 test using 57,170 shared SNPs in all AIL generations. We performed mega-analysis of F_{34} and F_{39-43} animals (n = 1,028) for body weight, coat color, and locomotor activity, the set of traits that were measured in the same way in both cohorts.

DISCUSSION

We used F_{34} and F_{39-43} generations of a LG/J x SM/J AIL to perform GWAS, SNP heritability estimates, genetic correlations, replication and mega-analysis. We had previously performed several GWAS using a sparse marker set in the F_{34} cohort. In this study we used a denser set of SNPs, obtained using GBS, to reanalyze the F_{34} cohort. We found 109 significant loci, 36 of which had not been identified in our prior studies using the sparse marker set. We used a new, previously unpublished F_{39-43} cohort for GWAS and showed that genetic correlations were high for the subset of traits that were measured in both cohorts. Despite this, we found that many loci were not replicated between cohorts, even when we used a relatively liberal definition of replication (p < 0.05). The failure to replicate some of our findings was not predicted by our power simulations. Therefore, we performed an analysis to determine whether Winner's Curse and study-specific heterogeneity could account for the lower than expected replication rate. Winner's Curse alone could not explain the failure to replicate. However, modeling both Winner's Curse and study-specific heterogeneity

better explained the observed replication rate. Finally, mega-analysis of the two cohorts allowed us to discover four additional loci. Taken together, our results provide a set of refined regions of association for numerous physiological and behavioral traits in multiple generations of AILs. These loci could serve as benchmarks for future GWAS results in intercross mouse lines. More broadly, this study illustrates the difficulty of replication even when using a highly controlled model system.

Previous publications from our lab used a sparse set of array genotypes for GWAS of various behavioral and physiological traits in 688 F_{34} AILs (Cheng et al. 2010; Lionikas et al. 2010; Samocha et al. 2010; Parker et al. 2011, 2014; Carroll et al. 2017; Hernandez Cordero et al. 2018; Gonzales et al. 2018). In this study we obtained a much denser marker set for 428 of the initial 688 AIL mice using GBS. The denser genotypes allowed us to identify most of the loci obtained using the sparse set, as well as many additional loci. For instance, using the sparse markers we identified a significant locus on chromosome 8 for locomotor day 2 activity that contained only one gene: *Csmd1* (CUB and sushi multiple domains 1). Gonzales et al. (Gonzales et al. 2018)

Phenotype	rG(s.e.)	SNP	Discovery set p (F$_{34}$ GBS)	-log10(p)	af	beta	se	Replication set p (F$_{3943}$ GBS replicate)	-log10(p)	af	beta	se
Body weight	0.711(0.25)*	chr4.66414508	8.58×10^{-8}	7.07	0.419	−0.25	0.05	3.55×10^{-3}	2.45	0.406	−0.13	0.04
		chr6.81405109	6.22×10^{-6}	5.21	0.497	0.21	0.05	3.52×10^{-2}	1.45	0.518	0.09	0.04
		chr14.79312393	7.45×10^{-6}	5.13	0.514	−0.20	0.04	2.37×10^{-2}	1.63	0.566	−0.10	0.04
Coat color, albino	0.967(0.04)*	chr7.87642045	5.00×10^{-106}	105.30	0.432	−0.58	0.02	1.59×10^{-162}	161.80	0.388	−0.61	0.02
Coat color, agouti	0.971(0.04)*	chr2.154464466	9.43×10^{-191}	190.03	0.129	0.94	0.01	5.7×10^{-93}	92.24	0.207	0.72	0.03
Locomotor test day 1, total distance traveled in 30min	0.968(0.24)*	chr19.21812298	3.98×10^{-7}	6.40	0.461	−0.36	0.07	4.55×10^{-1}	0.342	0.502	−0.05	0.06
Locomotor test day2, total distance traveled in 30min	0.988(0.19)*	chr8.17410225	5.65×10^{-6}	5.248	0.171	0.42	0.09	8.34×10^{-1}	0.079	0.202	0.02	0.08

Phenotype	rG(s.e.)	SNP	Discovery set p (F$_{3943}$ GBS)	-log10(p)	af	beta	se	Replication set p (F$_{34}$ GBS replicate)	-log10(p)	af	beta	se
Body weight	0.711(0.25)*	chr1.89192209	6.42×10^{-6}	5.19	0.22	0.22	0.05	5.16×10^{-2}	1.29	0.276	0.10	0.05
		chr14.82586326	1.48×10^{-6}	5.83	0.658	−0.22	0.04	3.08×10^{-5}	4.51	0.575	−0.19	0.05
Coat color, albino	0.967(0.04)*	chr7.87255156	3.37×10^{-166}	165.47	0.389	−0.62	0.02	7.80×10^{-97}	96.11	0.444	−0.57	0.02
Coat color, agouti	0.971(0.04)*	chr2.155091628	1.78×10^{-115}	114.75	0.218	0.74	0.02	1.51×10^{-185}	184.82	0.135	0.90	0.01
Locomotor test day 2, total distance traveled in 30min	0.988(0.19)*	chr15.67627183	3.33×10^{-6}	5.478	0.461	0.30	0.06	2.07×10^{-1}	0.683	0.522	−0.08	0.07

Discovery set	Replication set	Phenotype	Empirical replication rate	Predicted replication rate (**WC**)	Predicted replication rate (**WC+C**)
F_{34} GBS	F_{39-43} GBS	Body weight	0.009	1.000	0.044
		Coat color, albino	1.000	1.000	0.997
		Coat color, agouti	0.932	1.000	0.577
		Locomotor test day 1	0.000	1.000	0.028
		Locomotor test day 2	0.000	1.000	0.140
F_{39-43} GBS	F_{34} GBS	Body weight	0.297	1.000	0.071
		Coat color, albino	0.911	1.000	0.932
		Coat color, agouti	0.815	1.000	0.925
		Locomotor test day 2	0.000	1.000	0.053

replicated this finding in F_{50-56} AILs and identified a *cis*-eQTL mapped to the same region. *Csmd1* mutant mice showed increased locomotor activity compared to wild-type and heterozygous mice, indicating that *Csmd1* is likely a causal gene for locomotor and related traits (Gonzales *et al.* 2018). We replicated this locus in the analysis of the F_{34} cohort that used the denser marker set (Figure S8). We also replicated a locus on chromosome 17 for distance traveled in the periphery in the open field test (Figure 4; (Parker *et al.* 2014)), three loci on chromosomes 4, 6, and 14 for body weight (Figure S8; (Parker *et al.* 2011)), one locus on chromosome 7 for mean corpuscular hemoglobin concentrations (MCHC, complete blood count; Figure S8; (Bartnikas *et al.* 2012)), and numerous loci on chromosome 4, 6, 7, 8, and 11 for muscle weights (Figure S8; (Lionikas *et al.* 2010)). We noticed that even using original sparse markers, some previously published loci were not replicated in the current GWAS. The most likely explanation is that we had only 428 of the 688 mice used in the previous publications.

QTL mapping studies have traditionally used a 1.0~2.0 LOD support interval to approximate the size of the association region (see (Cervino *et al.* 2005; Logan *et al.* 2013)). The LOD support interval, proposed by Conneally *et al.* (Conneally *et al.* 1985) and Lander and Botstein (Lander and Botstein 1989), is a simple confidence interval method involving converting the p-value of the peak locus into a LOD score, subtracting "drop size" from the peak locus LOD score, and finding the two physical positions to the left and to the right of the peak locus location that correspond to the subtracted LOD score. Although Mangin *et al.* (Mangin *et al.* 1994) showed via simulation that the boundaries of LOD support intervals depend on effect size, others observed that a 1.0 ~2.0 LOD support interval accurately captures ~95% coverage of the true location of the loci when using a dense set of markers (Lander and Botstein 1989; Dupuis and Siegmund 1999; Manichaikul *et al.* 2006). In the present study, we considered using LOD support intervals but found that the sparse array SNPs produced misleadingly large support intervals. Various methods have been proposed for calculating confidence intervals in analogous situations (*e.g.*, (Baud *et al.* 2013; Nicod *et al.* 2016)). We performed credible set analysis and compared LocusZoom plots of the same locus region between array SNPs and the GBS SNPs (Figure S8; (Pruim *et al.* 2010)). For example, the benefit of the denser SNP coverage is easily observed in the locus on chromosome 7 (array lead SNP chr7:44560350; GBS lead SNP chr7:44630890) for the complete blood count trait "retic parameters cell hemoglobin concentration mean, repeat"; denser SNPs delineate the start and the end of an association block much

more clearly. Thus, there are advantages of dense SNP sets that go beyond the ability to discover additional loci.

LD in the LG/J x SM/J AIL mice is more extensive than in the Diversity Outbred mice and Carworth Farms White mice (Parker *et al.* 2016). Some of the loci that we identified are relatively broad, making it difficult to infer which genes are responsible for the association. We focused on loci that contained five or fewer genes (Table S6). We highlight a few genes that are supported by the existing literature for their role in the corresponding traits. The lead SNP at chr1:77255381 is associated with tibia length in F_{34} AILs (Table S6; Figure S8). One gene at this locus, *EphA4*, codes for a receptor for membrane-bound ephrins. EPHA4 plays an important role in the activation of the tyrosine kinase JAK2 and the signal transducer and transcriptional activator STAT5B in muscle, promoting the synthesis of insulin-like growth factor 1 (IGF-1) (Lai *et al.* 2004; Hyun 2013; Sawada *et al.* 2017). Mice with mutated *EphA4* shows significant defect in body growth (Hyun 2013). Curiously, another gene at this locus, *Pax3*, has been shown as a transcription factor expressed in resident muscle progenitor cells and is essential for the formation of skeletal muscle in mice (Relaix *et al.* 2006). It is possible that both *EphA4* and *Pax3* are associated with the trait tibia length because they are both involved in organismal growth. Another region of interest is the locus at chr4:66866758, which is associated with body weight (Table S6; Table S13). The lead SNP is immediately upstream of *Tlr4*, Toll-like receptor 4, which recognizes Gram-negative bacteria by its cell wall component, lipopolysaccharide (Hoshino *et al.* 1999; Takeuchi *et al.* 1999). TLR4 responds to the high circulating level of fatty acids and induces inflammatory signaling, which leads to insulin resistance (Shi *et al.* 2006). Kim *et al.* showed TLR-4-deficient mice were protected from the increase in proinflammatory cytokine level and gained less weight than wild-type mice when fed on high fat diet (Kim *et al.* 2012). The association between *Tlr4* and body weight in the AILs corroborates these findings.

We considered both the F_{34} and the F_{39-43} as both "discovery" and "replication" cohorts. Significant loci for coat color, which are monogenic and served as positive controls, were replicated, between the two cohorts, as expected. One locus for body weight was replicated (p < 0.05) between F_{34} and F_{39-43}. However, the loci for locomotor activity were not replicated. Power analyses predicted a much higher rate of replication, which led us to conduct additional analyses to better understand the lower than expected rate of replication.

First, we used a newly introduced method to determine whether Winner's Curse (Zöllner and Pritchard 2007)) which has also been

termed the Beavis Effect (Beavis et al. 1991, 1994; Xu 2003; King and Long 2017; Keele et al. 2019; Paterson 2019) could account for the lower than expected rate of replication. Beavis' original report described a lack of replication of QTL for agronomic traits between small populations of maize (Beavis et al. 1991). Using progeny sizes ranging from 100 to 1000, Beavis simulated interval mapping to evaluate the accuracy of the estimates of phenotypic variance explained at the statistically significant QTL (Beavis et al. 1994; Xu 2003; Paterson 2019). Simulations showed that progeny sizes greatly influenced the estimates of phenotypic variance explained; smaller progeny sizes (n = 100) generated highly overestimated estimates of phenotypic variances, whereas larger progeny sizes (n = 1000) generated estimates of phenotypic variances similar to the actual value (Xu 2003; Paterson 2019). King and Long (King and Long 2017) further examined the Beavis Effect in the next-generation mapping populations in *Drosophila melanogaster*. The authors found that sample size was the major determinant for the overestimation of phenotypic variance explained at the significant QTL in both the GWAS-based Drosophila Genetic Reference Panel (DGRP) and the multi-parental Drosophila Synthetic Population Resource (DSPR). When sample size remained constant and the true phenotypic variance explained at the significant QTL was small, the estimation bias was more pronounced. In contrast, estimates for the phenotypic variance explained at all simulated QTL, significant or not, were generally centered at the true values. In an analogous study of power and replication in Collaborative Cross mice, Keele et al. (Keele et al. 2019) found that the Beavis Effect was most striking when the number of strains and true effect size of the QTL were small. This estimation bias indicates that mapping statistically significant QTL across experiments, populations, and panels can be problematic (Macdonald and Long 2004; Gruber et al. 2007; Najarro et al. 2015). The analyses we performed indicated that Winner's Curse alone could not explain the lack of replication, but a model that also included study-specific heterogeneity could.

Our analysis does cannot explain the source of the study-specific heterogeneity. Possible sources of confounding could include maternal effects, which could differentiate the F_{34} cohort and the F_{39-43} cohort because F_{33} animals were transported to the University of Chicago from Washington University in St. Louis. In contrast, the breeder animals of the F_{39-43} cohort have already acclimated to the environment for multiple generations. Another possible source of confounding is that the phenotyping of the F_{39-43} occurred over five generations (more than a year) during which time numerous environmental factors may have changed (*e.g.*, several technicians performed the data collection). Such factors are known to be an important potential source of confounding; (Falconer 1960; Lynch and Walsh 1996; Crabbe et al. 1999; Mhyre et al. 2005; Visscher et al. 2008; Zaitlen and Kraft 2012; Sorge et al. 2014). Our analyses did not correct for the fact that six phenotypes were examined, thus somewhat increasing the chances that at least one of our significant associations could have been a false positive that would not be expected to replicate.

Interestingly, we found that the genetic correlations between the discovery and replication samples were relatively high for all traits; however, some traits replicated well and others replicated poorly. Our subsequent analysis showed that study-specific heterogeneity was low for the coat color traits, but higher for the body weight and locomotor traits. This makes an important point, namely that a strong genetic correlation can exists in the presence or absence of study-specific heterogeneity. Finally, it was notable that replication (at $p < 0.05$) was relatively successful for body weight, despite the significant evidence of study-specific heterogeneity and low predicted replication (Table 2). Power analyses predicted that the body weight loci should replicate at the genome-wide significance threshold, whereas we only

observed replication when at the less stringent $p < 0.05$ level (Table 1). The lack of replication at the genome-wide significance threshold for the body weight phenotype was likely due to study-specific heterogeneity due to confounding that was not accounted for in the power analyses. In Table 2, "predicted replication" refers to replication using a Bonferroni significance threshold that accounts for the number of significant SNPs in the discovery study. The low predicted replication rate under the **WC+C** model for the body weight phenotype is consistent with the low replication (genome-wide) reported in Table 1. Thus, both body weight and locomotor traits were strongly impacted by study specific confounding; however, nominal replication was still possible for body weight but not for the locomotor traits.

Finally, we performed a mega-analysis using F_{34} and F_{39-43} AIL mice. The combined dataset allowed us to identify four novel genome-wide significant associations that were not detected in either the F_{34} or the F_{39-43} cohorts presumably because of the increased sample size in the mega-analysis (Visscher et al. 2017). As is true for all GWAS, the loci identified in the mega-analysis could be false positives.

In addition to performing many GWAS and related analyses that led to the identification of dozens of novel loci for locomotor activity, open field test, fear conditioning, light dark test for anxiety, complete blood count, iron content in liver and spleen, and muscle weight, our study also tested our expectations about replication of GWAS findings. We did not obtain the expected rate of replication. We used a method that can distinguish between Winner's Curse and study-specific heterogeneity to show that the lower than expected rate of replication was due to study-specific heterogeneity. This indicates that study-specific heterogeneity can have a major impact of replication even when in a model system when a genetically identical population is tested under conditions that are designed to be as similar as possible.

LITERATURE CITED

Abney, M., 2008 Identity-by-descent estimation and mapping of qualitative traits in large, complex pedigrees. Genetics 179: 1577–1590. https://doi.org/10.1534/genetics.108.089912

Bartnikas, T. B., C. C. Parker, R. Cheng, D. R. Campagna, J. E. Lim et al., 2012 QTLs for murine red blood cell parameters in LG/J and SM/J F2 and advanced intercross lines. Mamm. Genome 23: 356–366. https://doi.org/10.1007/s00335-012-9393-3

Baud, A., V. Guryev, O. Hummel, M. Johannesson, R. Hermsen et al., 2014 Genomes and phenomes of a population of outbred rats and its

progenitors. Sci. Data 1: 140011. Erratum: 140016. https://doi.org/10.1038/sdata.2014.11

Beavis, W. D., D. Grant, M. Albertsen, and R. Fincher, 1991 Quantitative trait loci for plant height in four maize populations and their associations with qualitative genetic loci. Theor. Appl. Genet. 83: 141–145. https://doi.org/10.1007/BF00226242

Beavis, W. D., O. S. Smith, D. Grant, and R. Fincher, 1994 Identification of Quantitative Trait Loci Using a Small Sample of Topcrossed and F4 Progeny from Maize. Crop Sci. 34: 882–896. https://doi.org/10.2135/cropsci1994.0011183X003400040010x

Besnier, F., P. Wahlberg, L. Rönnegård, W. Ek, L. Andersson et al., 2011 Fine mapping and replication of QTL in outbred chicken advanced intercross lines. Genet. Sel. Evol. 43: 3. https://doi.org/10.1186/1297-9686-43-3

Browning, S. R., and B. L. Browning, 2007 Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81: 1084–1097. https://doi.org/10.1086/521987

Carbonetto, P., R. Cheng, J. P. Gyekis, C. C. Parker, D. A. Blizard et al., 2014 Discovery and refinement of muscle weight QTLs in B6 × D2 advanced intercross mice. Physiol. Genomics 46: 571–582. https://doi.org/10.1152/physiolgenomics.00055.2014

Carroll, A. M., R. Cheng, E. S. R. Collie-Duguid, C. Meharg, M. E. Scholz et al., 2017 Fine-mapping of genes determining extrafusal fiber properties in murine soleus muscle. Physiol. Genomics 49: 141–150. https://doi.org/10.1152/physiolgenomics.00092.2016

Cervino, A. C., G. Li, S. Edwards, J. Zhu, C. Laurie et al., 2005 Integrating QTL and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels. Genomics 86: 505–517. https://doi.org/10.1016/j.ygeno.2005.07.010

Cheng, R., M. Abney, A. A. Palmer, and A. D. Skol, 2011 QTLRel: An R package for genome-wide association studies in which relatedness is a concern. BMC Genet. 12: 66. https://doi.org/10.1186/1471-2156-12-66

Cheng, R., J. E. Lim, K. E. Samocha, G. Sokoloff, M. Abney et al., 2010 Genome-wide association studies and the problem of relatedness among advanced intercross lines and other highly recombinant populations. Genetics 185: 1033–1044. https://doi.org/10.1534/genetics.110.116863

Cheng, R., C. C. Parker, M. Abney, and A. A. Palmer, 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. https://doi.org/10.1534/g3.113.007948

Chesler, E. J., 2014 Out of the bottleneck: The Diversity Outcross and Collaborative Cross mouse populations in behavioral genetics research. Mamm. Genome 25: 3–11. https://doi.org/10.1007/s00335-013-9492-9

Churchill, G. A., D. M. Gatti, S. C. Munger, and K. L. Svenson, 2012 The diversity outbred mouse population. Mamm. Genome 23: 713–718. https://doi.org/10.1007/s00335-012-9414-2

Cockram, J., and I. Mackay, 2018 Genetic Mapping Populations for Conducting High-Resolution Trait Mapping in Plants. https://doi.org/10.1007/10_2017_48

Conneally, P. M., J. H. Edwards, K. K. Kidd, J.-M. Lalouel, N. E. Morton et al., 1985 Report of the committee on methods of linkage analysis and reporting. Cytogenet. Genome Res. 40: 356–359. https://doi.org/10.1159/000132186

Collaborative Cross Consortium, 2012 The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics 190: 389–401. https://doi.org/10.1534/genetics.111.132639

Coyner, J., J. L. McGuire, C. C. Parker, R. J. Ursano, A. A. Palmer et al., 2014 Mice selectively bred for High and Low fear behavior show differences in the number of pMAPK (p44/42 ERK) expressing neurons in lateral amygdala following Pavlovian fear conditioning. Neurobiol. Learn. Mem. 112: 195–203. https://doi.org/10.1016/j.nlm.2013.06.010

Crabbe, J. C., D. Wahlsten, and B. C. Dudek, 1999 Genetics of mouse behavior: Interactions with laboratory environment. Science 284: 1670–1672. https://doi.org/10.1126/science.284.5420.1670

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

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen et al., 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12: 499–510. https://doi.org/10.1038/nrg3012

Demarest, K., J. Koyner, J. McCaughran, L. Cipp, and R. Hitzemann, 2001 Further characterization and high-resolution mapping of quantitative trait loci for ethanol-induced locomotor activity. Behav. Genet. 31: 79–91. https://doi.org/10.1023/A:1010261909853

Diouf, I. A., L. Derivot, F. Bitton, L. Pascual, and M. Causse, 2018 Water Deficit and Salinity Stress Reveal Many Specific QTL for Plant Growth and Fruit Quality Traits in Tomato. Front. Plant Sci. 9: 279. https://doi.org/10.3389/fpls.2018.00279

Doitsidou, M., S. Jarriault, and R. J. Poole, 2016 Next-generation sequencing-based approaches for mutation mapping and identification in Caenorhabditis elegans. Genetics 204: 451–474. https://doi.org/10.1534/genetics.115.186197

Dupuis, J., and D. Siegmund, 1999 Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151: 373–386.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto et al., 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379. https://doi.org/10.1371/journal.pone.0019379

Falconer, D. S., (1960) Introduction to quantitative genetics. Introduction to Quantitative Genetics. https://www.cabdirect.org/cabdirect/abstract/19601603365

Fitzpatrick, C. J., S. Gopalakrishnan, E. S. Cogan, L. M. Yager, P. J. Meyer et al., 2013 Variation in the form of Pavlovian conditioned approach behavior among outbred male Sprague-Dawley rats from different vendors and colonies: Sign-tracking vs. goal-tracking. PLoS One 8: e75042. https://doi.org/10.1371/journal.pone.0075042

Gardenghi, S., M. F. Marongiu, P. Ramos, E. Guy, L. Breda et al., 2007 Ineffective erythropoiesis in β-thalassemia is characterized by increased iron absorption mediated by down-regulation of hepcidin and up-regulation of ferroportin. Blood 109: 5027–5035. https://doi.org/10.1182/blood-2006-09-048868

Gatti, D. M., K. L. Svenson, A. Shabalin, L.-Y. Wu, W. Valdar et al., 2014 Quantitative trait locus mapping methods for diversity outbred mice. G3: Genes, Genomes. Genetics 4: 1623–1633.

Ghazalpour, A., S. Doss, H. Kang, C. Farber, P.-Z. Wen et al., 2008 High-resolution mapping of gene expression using association in an outbred mouse stock. PLoS Genet. 4: e1000149. https://doi.org/10.1371/journal.pgen.1000149

Gonzales, N. M., and A. A. Palmer, 2014 Fine-mapping QTLs in advanced intercross lines and other outbred populations. Mamm. Genome 25: 271–292. https://doi.org/10.1007/s00335-014-9523-1

Gonzales, N. M., J. Seo, A. I. Hernandez Cordero, C. L. St Pierre, J. S. Gregory et al., 2018 Genome wide association analysis in a mouse advanced intercross line. Nat. Commun. 9: 5162. https://doi.org/10.1038/s41467-018-07642-8

Gonzales, N. M., J. Seo, A. I. Hernandez-Cordero, C. L. S. Pierre, J. S. Gregory et al., 2017 Genome wide association study of behavioral, physiological and gene expression traits in a multigenerational mouse intercross. Nat Commun 9: 5162. https://doi.org/10.1038/s41467-018-07642-8

Graziano, J. H., R. W. Grady, and A. Cerami, 1974 The identification of 2, 3-dihydroxybenzoic acid as a potentially useful iron-chelating drug. J. Pharmacol. Exp. Ther. 190: 570–575.

Gruber, J. D., A. Genissel, S. J. Macdonald, and A. D. Long, 2007 How Repeatable Are Associations Between Polymorphisms in achaete–scute and Bristle Number Variation in Drosophila? Genetics 175: 1987–1997. https://doi.org/10.1534/genetics.106.067108

Han, B., H. M. Kang, and E. Eskin, 2009 Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. PLoS Genet. 5: e1000456. https://doi.org/10.1371/journal.pgen.1000456

Han, L., and M. Abney, 2011 Identity by descent estimation with dense genome-wide genotype data. Genet. Epidemiol. 35: 557–567. https://doi.org/10.1002/gepi.20606

Hernandez Cordero, A. I., P. Carbonetto, G. Riboni Verri, J. S. Gregory, D. J. Vandenbergh et al., 2018 Replication and discovery of musculoskeletal QTLs in LG/J and SM/J advanced intercross lines. Physiol. Rep. 6: e13561. https://doi.org/10.14814/phy2.13561

Hernandez Cordero, A. I., N. M. Gonzales, C. C. Parker, G. Sokolof, D. J. Vandenbergh et al., 2019 Genome-wide Associations Reveal Human-Mouse Genetic Convergence and Modifiers of Myogenesis, CPNE1 and STC2. Am. J. Hum. Genet. 105: 1222–1236. Erratum: 138. https://doi.org/10.1016/j.ajhg.2019.10.014

Hoshino, K., O. Takeuchi, T. Kawai, H. Sanjo, T. Ogawa et al., 1999 Cutting Edge: Toll-Like Receptor 4 (TLR4)-Deficient Mice Are Hyporesponsive to Lipopolysaccharide: Evidence for TLR4 as the Lps Gene Product. J. Immunol. 162: 3749–3752.

Hyun, S., 2013 Body size regulation and insulin-like growth factor signaling. Cell. Mol. Life Sci. 70: 2351–2365. https://doi.org/10.1007/s00018-013-1313-5

Johnsson, M., R. Henriksen, A. Höglund, J. Fogelholm, P. Jensen et al., 2018 Genetical genomics of growth in a chicken model. BMC Genomics 19: 72. https://doi.org/10.1186/s12864-018-4441-3

Joo, J. W. J., F. Hormozdiari, B. Han, and E. Eskin, 2016 Multiple testing correction in linear mixed models. Genome Biol. 17: 62. https://doi.org/10.1186/s13059-016-0903-6

Keele, G. R., W. L. Crouse, S. N. P. Kelada, and W. Valdar, 2019 Determinants of QTL Mapping Power in the Realized Collaborative Cross. G3: Genes, Genomes. Genetics 9: 1707–1727. https://doi.org/10.1534/g3.119.400194

Kim, K.-A., W. Gu, I.-A. Lee, E.-H. Joh, and D.-H. Kim, 2012 High Fat Diet-Induced Gut Microbiota Exacerbates Inflammation and Obesity in Mice via the TLR4 Signaling Pathway. PLoS One 7: e47713. https://doi.org/10.1371/journal.pone.0047713

King, E. G., and A. D. Long, 2017 The Beavis Effect in Next-Generation Mapping Panels in Drosophila melanogaster. G3: Genes, Genomes. Genetics 7: 1643–1652. https://doi.org/10.1534/g3.117.041426

King, E. G., Macdonald, S. J., and Long, A. D. (2012) Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. Genetics, 112: 935–949.

Kislukhin, G., E. G. King, K. N. Walters, S. J. Macdonald, and A. D. Long, 2013 The genetic architecture of methotrexate toxicity is similar in Drosophila melanogaster and humans. G3 (Bethesda) 3: 1301–1310.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen, 2014 ANGSD: Analysis of next generation sequencing data. BMC Bioinformatics 15: 356. https://doi.org/10.1186/s12859-014-0356-4

Lai, K.-O., Y. Chen, H.-M. Po, K.-C. Lok, K. Gong et al., 2004 Identification of the Jak/Stat Proteins as Novel Downstream Targets of EphA4 Signaling in Muscle: implications in the regulation of acetylcholinesterase expression. J. Biol. Chem. 279: 13383–13392. https://doi.org/10.1074/jbc.M313356200

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

Lionikas, A., R. Cheng, J. E. Lim, A. A. Palmer, and D. A. Blizard, 2010 Fine-mapping of muscle weight QTL in LG/J and SM/J intercrosses. Physiol. Genomics 42A: 33–38. https://doi.org/10.1152/physiolgenomics.00100.2010

Listgarten, J., C. Lippert, C. M. Kadie, R. I. Davidson, E. Eskin et al., 2012 Improved linear mixed models for genome-wide association studies. Nat. Methods 9: 525–526. https://doi.org/10.1038/nmeth.2037

Logan, R. W., R. F. Robledo, J. M. Recla, V. M. Philip, J. A. Bubier et al., 2013 High-precision genetic mapping of behavioral traits in the diversity outbred mouse population. Genes Brain Behav. 12: 424–437. https://doi.org/10.1111/gbb.12029

Lynch, M., and B. Walsh, 1996 Genetics and Analysis of Quantitative Traits. Sunderland, MA: Sinauer.

Macdonald, S. J., and A. D. Long, 2004 A Potential Regulatory Polymorphism Upstream of hairy Is Not Associated With Bristle Number

Variation in Wild-Caught Drosophila. Genetics 167: 2127–2131. https://doi.org/10.1534/genetics.104.026732

Mackay, T. F., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles et al., 2012 The Drosophila melanogaster genetic reference panel. Nature 482: 173–178. https://doi.org/10.1038/nature10811

Mangin, B., B. Goffinet, and A. Rebai, 1994 Constructing confidence intervals for QTL location. Genetics 138: 1301–1308.

Manichaikul, A., J. Dupuis, S. Sen, and K. W. Broman, 2006 Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. Genetics 174: 481–489. https://doi.org/10.1534/genetics.106.061549

Marriage, T. N., E. G. King, A. D. Long, and S. J. Macdonald, 2014 Fine-mapping nicotine resistance loci in Drosophila using a multiparent advanced generation inter-cross population. Genetics 198: 45–57. https://doi.org/10.1534/genetics.114.162107

Mhyre, T. R., E. J. Chesler, M. Thiruchelvam, C. Lungu, D. A. Cory-Slechta et al., 2005 Heritability, correlations and in silico mapping of locomotor behavior and neurochemistry in inbred strains of mice. Genes Brain Behav. 4: 209–228. https://doi.org/10.1111/j.1601-183X.2004.00102.x

Najarro, M. A., J. L. Hackett, B. R. Smith, C. A. Highfill, E. G. King et al., 2015 Identifying Loci Contributing to Natural Variation in Xenobiotic Resistance in Drosophila. PLoS Genet. 11: e1005663. Erratum: e1005830. https://doi.org/10.1371/journal.pgen.1005663

Nicod, J., R. W. Davies, N. Cai, C. Hassett, L. Goodstadt et al., 2016 Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. Nat. Genet. 48: 912–918. https://doi.org/10.1038/ng.3595

Nikolskiy, I., D. F. Conrad, S. Chun, J. C. Fay, J. M. Cheverud et al., 2015 Using whole-genome sequences of the LG/J and SM/J inbred mouse strains to prioritize quantitative trait genes and nucleotides. BMC Genomics 16: 415. https://doi.org/10.1186/s12864-015-1592-3

Noldus, L. P. J. J., A. J. Spink, and R. A. J. Tegelenbosch, 2001 EthoVision: A versatile video tracking system for automation of behavioral experiments. Behav. Res. Methods Instrum. Comput. 33: 398–414. https://doi.org/10.3758/BF03195394

Parker, C. C., P. Carbonetto, G. Sokoloff, Y. J. Park, M. Abney et al., 2014 High-resolution genetic mapping of complex traits from a combined analysis of F2 and advanced intercross mice. Genetics 198: 103–116. https://doi.org/10.1534/genetics.114.167056

Parker, C. C., R. Cheng, G. Sokoloff, J. E. Lim, A. D. Skol et al., 2011 Fine-mapping alleles for body weight in LG/J × SM/J F_2 and F(34) advanced intercross lines. Mamm. Genome 22: 563–571. https://doi.org/10.1007/s00335-011-9349-z

Parker, C. C., R. Cheng, G. Sokoloff, and A. A. Palmer, 2012 Genome-wide association for methamphetamine sensitivity in an advanced intercross mouse line. Genes Brain Behav. 11: 52–61. https://doi.org/10.1111/j.1601-183X.2011.00747.x

Parker, C. C., S. Gopalakrishnan, P. Carbonetto, N. M. Gonzales, E. Leung et al., 2016 Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice. Nat. Genet. 48: 919–926. https://doi.org/10.1038/ng.3609

Parker, C. C., and A. A. Palmer, 2011 Dark matter: Are mice the solution to missing heritability? Front. Genet. 2: 32. https://doi.org/10.3389/fgene.2011.00032

Paterson, A. H., 2019 Molecular Dissection of Complex Traits, CRC Press. Boca Raton, FL.

Pruim, R. J., R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines et al., 2010 LocusZoom: Regional visualization of genome-wide association scan results. Bioinformatics 26: 2336–2337. https://doi.org/10.1093/bioinformatics/btq419

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira et al., 2007 PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575. https://doi.org/10.1086/519795

Rat Genome Sequencing and Mapping Consortium; Baud, A., R. Hermsen, V. Guryev, P. Stridh, D. Graham et al., 2013 Combined sequence-based

and genetic mapping analysis of complex traits in outbred rats. Nat. Genet. 45: 767–775. https://doi.org/10.1038/ng.2644

Relaix, F., D. Montarras, S. Zaffran, B. Gayraud-Morel, D. Rocancourt et al., 2006 Pax3 and Pax7 have distinct and overlapping functions in adult muscle progenitor cells. J. Cell Biol. 172: 91–102. https://doi.org/10.1083/jcb.200508044

Rishmawi, L., J. Bühler, B. Jaegle, M. Hülskamp, and M. Koornneef, 2017 Quantitative trait loci controlling leaf venation in Arabidopsis. Plant Cell Environ. 40: 1429–1441. https://doi.org/10.1111/pce.12938

Samocha, K. E., J. E. Lim, R. Cheng, G. Sokoloff, and A. A. Palmer, 2010 Fine mapping of QTL for prepulse inhibition in LG/J and SM/J mice using F2 and advanced intercross lines. Genes Brain Behav. 9: 759–767. https://doi.org/10.1111/j.1601-183X.2010.00613.x

Sawada, T., D. Arai, X. Jing, M. Miyajima, S. J. Frank et al., 2017 Molecular interactions of EphA4, growth hormone receptor, Janus kinase 2, and signal transducer and activator of transcription 5B. PLoS One 12: e0180785. https://doi.org/10.1371/journal.pone.0180785

Shi, H., M. V. Kokoeva, K. Inouye, I. Tzameli, H. Yin et al., 2006 TLR4 links innate immunity and fatty acid–induced insulin resistance. J. Clin. Invest. 116: 3015–3025. https://doi.org/10.1172/JCI28898

Sittig, L. J., P. Carbonetto, K. A. Engel, K. S. Krauss, C. M. Barrios-Camacho et al., 2016 Genetic Background Limits Generalizability of Genotype-Phenotype Relationships. Neuron 91: 1253–1259. https://doi.org/10.1016/j.neuron.2016.08.013

Sorge, R. E., L. J. Martin, K. A. Isbester, S. G. Sotocinal, S. Rosen et al., 2014 Olfactory exposure to males, including men, causes stress and related analgesia in rodents. Nat. Methods 11: 629–632. https://doi.org/10.1038/nmeth.2935

Svenson, K. L., D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng et al., 2012 High-resolution genetic mapping using the Mouse Diversity outbred population. Genetics 190: 437–447. https://doi.org/10.1534/genetics.111.132597

Takeuchi, O., K. Hoshino, T. Kawai, H. Sanjo, H. Takada et al., 1999 Differential Roles of TLR2 and TLR4 in Recognition of Gram-Negative and Gram-Positive Bacterial Cell Wall Components. Immunity 11: 443–451. https://doi.org/10.1016/S1074-7613(00)80119-3

Talbot, C. J., A. Nicod, S. S. Cherny, D. W. Fulker, A. C. Collins et al., 1999 High-resolution mapping of quantitative trait loci in outbred mice. Nat. Genet. 21: 305–308. https://doi.org/10.1038/6825

Toker, L., M. Feng, and P. Pavlidis, 2016 Whose sample is it anyway? Widespread misannotation of samples in transcriptomics studies. F1000 Res. 5: 2103. https://doi.org/10.12688/f1000research.9471.1

Valdar, W., L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman et al., 2006 Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat. Genet. 38: 879–887. https://doi.org/10.1038/ng1840

Visscher, P. M., W. G. Hill, and N. R. Wray, 2008 Heritability in the genomics era—Concepts and misconceptions. Nat. Rev. Genet. 9: 255–266. https://doi.org/10.1038/nrg2322

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy et al., 2017 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. 101: 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Vonesch, S. C., D. Lamparter, T. F. Mackay, S. Bergmann, and E. Hafen, 2016 Genome-wide analysis reveals novel regulators of growth in Drosophila melanogaster. PLoS Genet. 12: e1005616. https://doi.org/10.1371/journal.pgen.1005616

Wellcome Trust Case Control Consortium; Maller, J. B., G. McVean, J. Byrnes, D. Vukcevic, K. Palin, 2012 Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat. Genet. 44: 1294–1301. https://doi.org/10.1038/ng.2435

Xu, S., 2003 Theoretical Basis of the Beavis Effect. Genetics 165: 2259–2268.

Yalcin, B., S. A. Willis-Owen, J. Fullerton, A. Meesaq, R. M. Deacon et al., 2004 Genetic dissection of a behavioral quantitative trait locus shows that Rgs2 modulates anxiety in mice. Nat. Genet. 36: 1197–1202. https://doi.org/10.1038/ng1450

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 GCTA: A tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88: 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price, 2014 Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. 46: 100–106. https://doi.org/10.1038/ng.2876

Yang, J., J. Zeng, M. E. Goddard, N. R. Wray, and P. M. Visscher, 2017 Concepts, estimation and interpretation of SNP-based heritability. Nat. Genet. 49: 1304–1310. https://doi.org/10.1038/ng.3941

Zaitlen, N., and P. Kraft, 2012 Heritability in the genome-wide association era. Hum. Genet. 131: 1655–1664. https://doi.org/10.1007/s00439-012-1199-6

Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44: 821–824. https://doi.org/10.1038/ng.2310

Zöllner, S., and J. K. Pritchard, 2007 Overcoming the Winner's Curse: Estimating Penetrance Parameters from Case-Control Data. Am. J. Hum. Genet. 80: 605–615. https://doi.org/10.1086/512821

Zou, J., J. Zhou, S. Faller, R. Brown, and E. Eskin, 2019 Accurate modeling of replication rates in genome-wide association studies by accounting for winner's curse and study-specific heterogeneity. bioRxiv 21: https://doi.org/10.1101/856898

Communicating editor: S. Macdonald