**Supplementary materials**

**Figure S1. Autosomal heterozygosity distribution in F34, F39-43 AILs.** Animals with excessive or insufficient heterozygosity (3 *s.d.* away from mean) were removed from further analysis. As controls, we have sequenced two F2s of LG and SM, four LG mice and four SM mice (see annotated data points with 1 and 0 heterozygosity).

**Figure S2. Kinship coefficients in F34 and F39-43 AILs calculated from pedigree against genetic relatedness matrix calculated using** IBDLD(L. Han and Abney 2011; Abney 2008)**.** Each circle represents a pair of animals, which their genetic kinship relatedness on the x-axis and pedigree kinship relatedness on the y-axis. Color signifies relatedness based on AIIL pedigree. Blue circles represent identical twins, red full siblings, yellow parent-offspring pairs, grey other relationships. Seven animal pairs that deviate from the pedigree relationship clusters were excluded (see black arrows).

**Figure S3. Heatmap showing F34 array and F34 GBS genotype concordance in percentages, using 66 shared SNPs.** “A” codes for the LG/J allele, and “B” codes for the SM/J allele. “AA” genotype concordance between array and GBS is 24.54%, “AB” 43.23%, “BB” 27.60%.

**Figure S4. LD decay in F34 array, F34 GBS, F39-43 GBS, and F34 and F39-43 GBS SNP sets.** Average LD (*r2*) was calculated using allele frequency matched SNPs (MAF difference < 0.05), as described in Parker et al. (Parker et al. 2016). We found that LD decay rates using F34 array, F34 GBS, F39-43 GBS, and F34 and F39-43 GBS genotypes were generally similar to one another.

**Figure S5. SNP heritability using F34 GBS and F34 array SNPs (slope=1).** SNP heritability was estimated using **GCTA-GREML** (Yang et al. 2011)**. All traits except the coat color trait, agouti, have similar SNP heritability estimates between F34 array and F34 GBS SNPs.**

**Figure S6. Power simulations for discovery SNPs in the replication set.** Power was simulated at both the genome-wide significance level for the cohort and the nominal p-value of 0.05. Each data point represents the estimated power at the simulated beta. The vertical dashed line in orange indicates the effect size of the discovery SNP.

**Figure S7.** **Modeling study-specific heterogeneity due to confounding is necessary for explaining observed patterns of replication.** Graphical interpretations of the models are shown for the body weight, albino coat color, agouti coat color, and locomotor activity on day 1 and day 2 phenotypes. Each point corresponds to one variant that was genome-wide significant in the discovery study (F34). The black points correspond to the variants that were replicated using a Bonferroni threshold in F39-43, while the grey points correspond to variants that were not replicated. The x-axis shows the summary statistic obtained from the discovery study, and the y-axis shows the summary statistic obtained from the replication study. The solid lines correspond to the expected value of the replication statistics after accounting for 1) Winner’s Curse (**WC**) or 2) Winner’s Curse and confounding (**WC+C**). Locomotor Activity on Day 2 only had two variants that passed the genome-wide threshold; as a result, the estimates of the parameters may not be accurate. For albino coat color, the expected values are identical under the **WC** and **WC+C** models due to low estimates of study-specific heterogeneity. The dashed lines correspond to the 95% confidence intervals on those estimates. For body weight and locomotor activity on day 1, the axes have been constrained to better visualize the points, and the confidence intervals are outside of the axes limits.

**Figure S8. LocusZoom for F34 array, F34 GBS, F39-43 GBS, and mega-analysis QTLs.** Using the standalone implementation of LocusZoom (Pruim et al. 2010), we plotted LocusZoom regions for all significant QTL in all cohorts. Purpletrack shows the credible set interval (*r2* threshold = 0.8, posterior probability threshold = 0.99).

**Table S1. List of phenotypes used in GWAS.** Additional information for each trait, including the generation of the cohort, sample size, sex, age, and covariates included in the LMM model, was included in this table as well as on GeneNetwork2 (<http://gn2.genenetwork.org/>).

**Table S2. SNP and individual QC filter table.** Numbers of animals and SNPs remained after each step of filtering are shown per GBS SNP set.

**Table S3. GBS SNPs in the F34 cohort with HWE p-values close to 1.0×10-6 cutoff threshold.** These SNPs are removed from QTL summary tables.

**Table S4. Effect of PLINK v1.9 clump-based pruning parameters on number of independent SNPs remained.** At all *r2* values examined, a sliding window size of 12150kb was the first smallest window that yield the most stringent number of clumped SNPs in both array and GBS GWAS.

**Table S5. Adjusted significance threshold for each SNP set and GWAS cohort.** We used MultiTrans to obtain a unique adjusted significance threshold for each SNP set in each animal cohort (Joo et al. 2016; B. Han et al. 2009).

**Table S6. Select lead SNPs with association regions containing less than 5 coding genes.** Credible set analysis was performed to define the boundaries of the locus (*r2* threshold = 0.8, posterior probability threshold = 0.99). Genes contained in and/or immediately downstream of the credible set interval were included as associated genes.

**Table S7. Lead QTL in F34 GBS and F34 array GWAS studies across phenotypes.** Significant SNPs are clumped using parameters *r2*=0.1, 12150kb.

**Table S8. F34 GBS and array SNP heritability estimates.** SNP heritability estimates are performed using the GCTA-GREML analysis(Yang et al. 2011)**.**

**Table S9. Lead QTL in the F39-43 cohort (N=600).** Significant SNPs are clumped using parameters *r2*=0.1, 12150kb.

**Table S10. Replication of significant SNPs between the F34 cohort and the F39-43 cohort.** We performed the credible set analysis to provide a QTL interval that delineates an association region. Using the credible set interval defined for a discovery SNP, we searched for the SNP with the lowest p-value within the credible set region in the replication set. "Top SNP in Credible Set" denotes the replication SNP with the lowest p-value within the credible set interval of the discovery SNP. We found that the p-value (and -log10(p) value) for the "Matching SNP" and the "Top SNP in Credible Set" in the replication cohort were highly similar. Our results show that the credible set interval of the discovery SNP defines the association region of the loci fairly accurately.

**Table S11. F34 and F39-43 genetic correlations in locomotor activity, coat color, and body weight.** Genetic correlations were performed using the GCTA bivariate GREML analysis(Yang et al. 2011)**.**

**Table S12. SNP-heritability comparison between the F34 and F39-43 cohorts.** GBSSNP heritability was consistently lower in the F39-43 compared to the F34 cohort (Figure 3).

**Table S13. Lead QTL in the combined F34 and F39-43 cohort (N=1028).** Significant SNPs are clumped using parameters *r2*=0.1, 12150kb.

**File S1. Code used to perform analyses in the present study**, including Beagle imputation, GWAS dosage preparation, GRM estimation, running GEMMA, LD-based clumping, credible set analysis, significance threshold estimation using MultiTrans and SLIDE, power analysis, heritability estimates, and genetic correlations.