Title

The utility of a closed breeding colony of Peromyscus leucopus for dissecting complex traits

Journal

Genetics, 221(1)

ISSN

0016-6731

Authors

Long, Phillip N
Cook, Vanessa J
Majumder, Arundhati
et al.

Publication Date

2022-05-05

DOI

10.1093/genetics/iyac026

Peer reviewed

OXFORD GENETICS

GENETICS, 2022, 221(1), iyac026
https://doi.org/10.1093/genetics/iyac026
Advance Access Publication Date: 10 February 2022
Investigation

The utility of a closed breeding colony of *Peromyscus leucopus* for dissecting complex traits

Phillip N. Long,[1] Vanessa J. Cook,[2] Arundhati Majumder,[1] Alan G. Barbour,[1,2] and Anthony D. Long[1,*]

[1]Department of Ecology and Evolutionary Biology, School of Biological Sciences, University of California Irvine, Irvine, CA 92697-2525, USA, and
[2]Departments of Microbiology & Molecular Genetics and Medicine, School of Medical Sciences, University of California Irvine, Irvine, CA 92687-2525, USA

*Corresponding author: Department of Ecology and Evolutionary Biology, School of Biological Sciences, University of California Irvine, Irvine, CA 92697-2525, USA. Email: tdlong@uci.edu

Abstract

Deermice of the genus *Peromyscus* are well suited for addressing several questions of biologist interest, including the genetic bases of longevity, behavior, physiology, adaptation, and their ability to serve as disease vectors. Here, we explore a diversity outbred approach for dissecting complex traits in *Peromyscus leucopus*, a nontraditional genetic model system. We take advantage of a closed colony of deer-mice founded from 38 individuals and subsequently maintained for ∼40–60 generations. From 405 low-pass short-read sequenced deermice we accurate impute genotypes at 16 million single nucleotide polymorphisms. Conditional on observed genotypes simulations were conducted in which three different sized quantitative trait loci contribute to a complex trait under three different genetic models. Using a stringent significance threshold power was modest, largely a function of the percent variation attributable to the simulated quantitative trait loci, with the underlying genetic model having only a subtle impact. We additionally simulated 2,000 pseudo-individuals, whose genotypes were consistent with those observed in the genotyped cohort and carried out additional power simulations. In experiments employing more than 1,000 mice power is high to detect quantitative trait loci contributing greater than 2.5% to a complex trait, with a localization ability of ∼100 kb. We finally carried out a Genome-Wide Association Study on two demonstration traits, bleeding time and body weight, and uncovered one significant region. Our work suggests that complex traits can be dissected in founders-unknown *P. leucopus* colony mice and similar colonies in other systems using easily obtained genotypes from low-pass sequencing.

Keywords: multiparent panel; diversity outbred; heterogeneous stock; power; QTL mapping; peromyscus; genotyping; GWAS

Introduction

Variation in complex genetic traits is due to the action of many genes as well as the environment. Despite complex genetic traits (e.g. risk of certain mental disorders, heart disease, stroke, and diabetes) accounting for the bulk of US health spending, in most cases we do not yet understand their precise genetic architecture. Over the last decade, human geneticists have employed a Genome-Wide Association Study (GWAS) approach to identify over 70 thousand factors contributing to complex traits (Manolio *et al.* 2009; Buniello *et al.* 2019), with the vast majority of identified factors having extremely subtle phenotypic effects (Boyle *et al.* 2017). In contrast, in major genetic model systems a MultiParent Population (MPP) approach has emerged, where quantitative trait loci are mapped in panels derived from multigeneration crosses between several inbred parental lines (de Koning and McIntyre 2017). In the classic MPP-approach to quantitative trait loci (QTL) mapping, the MPP population is sampled at an advanced generation and fully genotyped Recombinant Inbred Lines (RILs) are derived. MPP-RILs are available in mice, Drosophila, maize, Arabidopsis, and so forth, and have been successfully used to map hundreds of traits (Long *et al.* 2014; de Koning and McIntyre 2017). Despite the success of MPP-RIL approaches, the cost of maintaining several hundred RILs can be high, and RILs are both "lost" and accumulate deleterious mutations over time. In contrast, as the cost of genotyping via short read sequencing continues to plummet we are increasingly seeing the growth of a "Diversity Outbred"/"Heterogeneous Stock" (DO/HS) paradigm where advanced generation MPPs individuals are directly used to map traits (Mott *et al.* 2000; Macdonald and Long 2007; Svenson *et al.* 2012). Indeed there are now multiple HS populations maintained in mice (Svenson *et al.* 2012; Woods and Mott 2017) and rats (Hansen and Spuhler 1984) used explicitly for this purpose.

Although mice and rats have excellent DO/HS populations there are many complex traits appropriately studied in other rodent systems. A rodent system of particular significance are "mice" of the *Peromyscus* genus. Although called mice, *Peromyscus* spp. are cricetine rodents, which include hamsters, and are only distantly related to the major murid genetic model systems. We henceforth refer to *Peromyscus* as "deermice" to avoid this potential confusion. Deermice are the one of the most abundant mammals in North America, and have enjoyed success as a model system for understanding the genetics of adaptation to high altitude, changes in metabolism and immune function during adaptation to urban environments, the genetics of longevity and life history, the genetics of behavioral traits (such as parental care

Received: December 14, 2021. **Accepted:** February 01, 2022

and nest building), and adaptive spatial changes in coat coloration to avoid predation among other traits (beautifully reviewed in Bedford and Hoekstra 2015). Different deermice species are also major reservoirs for several tick-borne diseases including Lyme disease, *Borrelia miyamotoi* relapsing fever, the malaria-line protozoan disease babesiosis, and hantavirus (Barbour 2017). The role of *Peromyscus leucopus* as the likely primary reservoir for the bacteria that causes Lyme disease (*Borrelia burgdorferi*) and several other tick-borne diseases is analogous to that of bats as reservoirs for SARS coronaviruses and Ebola virus. In fact, *P. leucopus*'s role as the primary reservoir for the bacteria that causes Lyme disease has led to proposals that it be the first mammal considered for natural release gene drive experiments in North America (Najjar *et al.* 2017).

Our previous effort to create infrastructure to strengthen *P. leucopus*'s role as an emerging model system for the study of infectious and other diseases produced a chromosome-length scaffolded 2.45 Gb genome assembly, demonstrated how this annotated assembly can empower RNAseq experiments, and identified ∼42 million intermediate frequency single nucleotide polymorphisms (SNPs) (Long *et al.* 2019). Despite having an annotated University of California Santa Cruz (UCSC) browsable genome, *P. leucopus* does not have a purpose-constructed HS/DO population that can be used for dissecting complex traits like those available in mice and rats. *Peromyscus leucopus* does have a closed breeding colony founded from 1982 to 1985 from 38 wild-caught deermice (the "*P. leucopus* White-footed mouse LL Stock") that shows great potential for the dissection of complex traits. We speculated that this colony, by virtue of its being closed for ∼60 generations since founding, could have many of the properties of the MPPs widely used in model systems. Importantly, we expect the colony to display increased levels of LD relative to natural populations, where useful LD only extends over a few hundred base-pairs (Long *et al.* 2019), and decreased levels of nucleotide variation. These features of the colony should allow missing genotypes to be imputed from low pass sequencing data and genome-wide association scans effectively carried out. Unlike the mouse and rat HS/DO populations, but like other outbred stocks that may exist for other emerging model species, the 38 wild-caught deermice "founders" of the *P. leucopus* LL colony are unknown. We show this lack of founder information may minimally impact on our ability to impute SNP genotypes and carry out a GWAS. Instead, the main limitation of having "unknown founders" is a reduced ability to identify candidate causative variants (based on the pattern of SNPs private to certain founders) after traits are mapped.

Here, we carry out ∼1X per animal low pass sequencing on 405 *P. leucopus* LL colony deermice and use `Stitch` (Davies *et al.* 2016) to impute SNPs (and haplotypes) for each individual. We validate imputed genotypes using RNAseq data obtained from a subset of 5 genotyped deermice and show imputed SNP genotypes are generally of high quality. We then simulate QTL contributing 1%, 2.5%, or 5% to variation in a complex trait under 3 different genetic models and examine the power of both a marker- and haplotype-based test to detect phenotype-genotype associations. Using a threshold for genome-wide testing that holds the false positive rate at less than 5% per genome scan and a study employing 348 deermice we show that the power to detect a simulated QTL is ∼1–50% largely irrespective of the underlying genetic model simulated, but dependent on the statistical test and the contribution of the simulated QTL to total phenotypic variation. Despite modest power at this sample size, we consistently observe a strong linkage signal and localized peaks at the locations of simulated QTL, suggesting larger sample sizes will greatly increase power, in a manner consistent with that observed in mouse and rat HS/DO studies (c.f. Gatti *et al.* 2014). We modify the `Stitch` machinery to simulate up to 2,000 pseudo-individuals whose genomes are conditional on the genotyped animals, show the power to map QTL can be high given a study consisting of >1,000 colony deermice, and further show that localization ability can exceed 100 kb. We finally carry out a GWAS study for 2 demonstration traits: the time between when a tail is clipped to obtain a small amount of tissue for genotyping and the associated wound cauterizes and log transformed body weight. We observe a single region exceeding the threshold for significance for body weight.

Materials and methods

Animals

Adult female and male outbred *P. leucopus* of the LL stock were obtained from the Peromyscus Genetic Stock Center (PGSC) of the University of South Carolina (Peromyscus Genetic Stock Center, 2017). The LL stock colony was founded with 38 animals captured near Linville, NC between 1982 and 1985 and has been closed since 1985. Sib-matings are proscribed, and complete pedigree records are kept. Animals of the LL stock have mitochondria with the same genome sequence (Barbour *et al.* 2019). The gastrointestinal microbiota of the colony animals have been characterized (Milovic *et al.* 2020). Animals were maintained in the AAALAC-accredited U.C. Irvine vivarium with 2–5 animals per cage according to sex and on 12 h light–12 h dark lighting schedule, temperature of 21–23°C, humidity of 30–70%, water ad libitum, and a diet of 8604 Teklad Rodent (Harlan Laboratories). The study was carried out in accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. University of California Irvine protocol AUP-18-020 was approved by the Institutional Animal Care and Use Committee (IACUC).

Bleeding time assay

The method was a modification of that of Broze *et al.* (2001). The animals were lightly anesthetized with 3% isoflurane by inhalation with 2 L/min flow of oxygen in a small animal veterinary induction chamber. A sterilized, small animal nail clipper (Conair PRO small; item PGRDNCS) equipped with a guard set at 2 mm was used to sever the tail's tip, and the timer was started. The exposed tissue was briefly touched every 0.5 min with Whatman No. 2 filter paper. The recorded bleeding time was the number of minutes in half minute intervals until further bleeding of the exposed tail tip ceased for at least 0.5 min after the last touch of filter paper. To prevent further bleeding after the animal was returned to its cage Blood Stop Powder (Durvet) was applied. Bleeding times of less than 1.0 min were not observed. The range of bleeding times was 1.0 to 13.0 min. For 103 animals for which the bleeding time was repeated 2–4 days later the coefficient of determination (R^2) was 0.87. Of the 103, for only 5 (4.9%) was the difference between the 2 bleeding determinations more than 1.0 min.

DNA extractions and short read libraries

The 2 mm sample of the excised terminal tail tissue was subjected to extraction with Qiagen's DNeasy Blood and Tissue Kit. The tissue was first placed in a 2 ml microcentrifuge tube, to which were then added 180 µl of the kit's Buffer ATL and 30 µl of Proteinase K (Qiagen) at 20 mg/ml. The tube was vortexed and

then placed on a shaker at 56°C at 200–220 rpm for up to 12 h until the skin tissue had dissolved. Following this step, the procedure then followed the manufacturer's instructions. DNA concentration was determined using a Qubit 2.0 fluorometer with the Qubit dsDNA HS Assay Kit. Short read libraries were prepared in 96-well plates using 1/5th size Illumina Nextera FLEX chemistry reactions, 50 ng of gDNA per mouse, and a custom set of 96 unique barcode pairs. We follow the Illumina protocol through the 12-cycle PCR amplification of tagged products, but then post PCR proteinase-K treat each of the 96 samples (to remove polymerase activity), create a 96-plex pool using 2 µl of each sample, and then clean-up a 45 µl aliquot of the 96-plex reaction following the FLEX protocol. Each 96-plex reaction was run as a PE100 over 3 HiSEQ4000 lanes, and deplexed using the Illumina software.

Processing of raw sequencing data

Raw fastq files for 405 animals were aligned to the reference genome using `bwa mem` v.0.7.8 (Li 2013) and default settings, with PCR duplicates removed and individuals having less than 0.1X coverage excluded from further consideration. We then used `Stitch` v.1.6.6 (Davies *et al.* 2016) to impute genotypes for each of the animals with default settings, and `"-K=8 -nGen=60 -output_haplotype_dosages=TRUE."` `Stitch` takes as input the bam files from the alignment step and a list of SNPs at which imputation takes place, here 17.75 M biallelic nonrepeat-overlapping SNPs, identified from a previous study that carried out sequencing of 36 colony individuals (Long *et al.* 2019). Only SNPs seen in at least 2 of the 36 colony animals were considered, thus the SNPs of this study are a subset of all SNPs in the colony, with a bias toward more common variants. The `nGen` parameter reflects the number of generations since colony founding (imputation does not seem very sensitive to this parameter). K is the number of founder haplotypes in the population. Although the number of founder haplotypes in the colony is 76, which is much greater than the employed $K=8$, increasing (or decreasing) K greatly beyond 8 seems to result in poorer quality imputations, so we adopted this number. We further show in the results that very few regions of the genome appear to segregate more than 8 common haplotypes. `Stitch` was initially run on all 405 low-pass sequenced individuals (to increase imputation accuracy), although the simulations of this work focus on a subset of 348 animals for which we measured a quantitative trait. `Stitch` outputs imputed genotypes and haplotype dosages as a vcf file. Genotype and haplotype dosages were extracted per chromosome using the following `bcftools` v.1.10.2 (Danecek *et al.* 2021) queries: `"%POS [%DS{0}]\n"` for genotypes and `"%POS [%HD{0} %HD{1} %HD{2} %HD{3} %HD{4} %HD{5} %HD{6} %HD{7}]\n"` for haplotypes, and wrangled into alternate formats using custom python scripts (see software). We finally excluded SNPs having a MAF <1% or an info score <0.4 in the 348 deermice.

Validation of Stitch genotype calls

`Stitch` outputs a "dosage" for each imputed SNP, where dosage is a number between 0 and 2 that estimates the number of nonreference alleles per individual. To validate `Stitch` dosage calls, we took advantage of 5 animals for which RNAseq data from spleens was available. RNAseq data from these animals was aligned to the genome using `hisat2` v.2.2.1 (Kim *et al.* 2019) and SNPs were called using `GATK` (version 4.1.9.0; McKenna *et al.* 2010). SNPs in the vcf file were filtered using the following `vcftools` v.1.10.2 (Danecek *et al.* 2011) switches `"--min-alleles 2 --max-alleles 2 --min-meanDP 20 --max-missing 1 --maf 0.15 --remove-indels --minQ 30"` which resulted in 110,188 SNP calls. These SNPs have an average coverage of 93X and as a result we expect the calls to be quite accurate in most cases. Although factors such as allelic imbalance or read mapping biases could result in a small fraction of the high-quality SNP calls obtained from RNAseq being incorrect, such errors will result in under-estimates of the accuracy of our SNP imputation approach. We compared the high-quality SNP calls from RNAseq to imputed `Stitch` dosages to calculate either the per SNP absolute error or the squared correlation coefficient (R^2) between estimates as a function of per sample gDNA short-read low-pass sequence coverage.

Genetic models

To assess the power and localization ability of experiments treating *Peromyscus* colony deermice like DO/HS mice or rats we simulated mapping experiments where a SNP or SNPs contributed to variation in a complex trait. As we do not know the true genetic architecture of complex traits, we explore 3 alternative models to ensure any statistical test is robust to model assumptions. Under all models a single "gene" contributed 1%, 2.5%, or 5% to variation in a complex trait, random Gaussian environmental variation contributed 50%, and polygenic background variance contributed 49%, 47.5%, or 45% to a complex trait. The first model assumes single causative SNP, the second assumes 10 SNPs in a 100 kb window are causative, and the third assumes all SNPs in a 100 kb window are causative. The latter 2 models attempt to simulate a causative "gene" contributing a significant proportion of variation to a complex trait, with variation at that gene due to several causative mutations. These multiple causative variants in a single gene simulations are consistent with models of mutation selection balance maintaining variation in complex traits (Pritchard and Cox 2002; Thornton *et al.* 2013). Polygenic background was simulated by randomly choosing every 400th SNP in the genome, adding dosage values within individuals, and re-scaling the resulting sums to their target variances. We simulated 2X250 different causative "genes," with each gene centered on the uniform random location of 250 SNPs (minor allele count greater than two) located on either Chromosome 23 or 19. Conditioning on a minor allele count of greater than 2 avoids studying private alleles of large effect whose imputation might be suspect. In the case of the single SNP causative model, the dosage associated with each causative SNP was rescaled to account for 1%, 2.5%, or 5% of variation in a complex trait. For the 10 and all SNP models, we defined a 100 kb region centered on each of the 250 SNPs (or "gene midpoints") and further considered 10 randomly chosen SNPs or all SNPs in the region. In the case of the 10 SNP model, we added the dosages and rescaled so that they accounted for 1%, 2.5%, or 5% of variation. In the case of the all SNP model, we scaled the dosages to have unit variance before adding them, and then rescaled so that they accounted for 1%, 2.5%, or 5% of variation in the trait. Thus, under the all SNP model, rare SNPs effectively have larger effect sizes, whereas under the 10-SNP model, intermediate frequency alleles had larger effects. We simulate the single SNP causative model for historical reasons (this model has been widely simulated), although hundreds of human GWAS studies suggest that single intermediate frequency SNPs accounting for more than 1% of the variation in a complex trait are extremely uncommon.

We simulate causative loci on either Chromosome 23 or 19, but we carry out genome-wide scans excluding markers on Chromosome 19. In the case of a causative "gene" or locus on Chromosome 23 and a genome-wide (less Chromosome 19) scan, hits are used to estimate the true positive rate of an experiment.

Whereas for causative loci on Chromosome 19 and the same genome-wide scan hits are used to estimate the false-positive rate and establish a threshold for statistical significance.

Kinship matrix

The individuals of this study are from a closed colony. Although the colony employs a breeding design to minimize inbreeding, mice can be closely related and the genetic constitution of the colony is slowly changing over time. It is common in such situations to employ a kinship matrix and mixed statistical models that incorporate a kinship matrix. We explored several methods for estimating a kinship matrix, including methods employing SNPs vs founder haplotype dosages. Different methods tend to produce similar, but not identical relatedness matrices, as assessed by examining the correlation between off-diagonal elements of the matrix. The kinship matrix we ultimately employed chose every 631st imputed SNP in the genome (with 631 chosen so SNPs used to construct the kinship matrix are "out of phase" with those used to simulate background polygenic variation), hard-encoded the dosage estimates to a genotype by rounding to the nearest whole number, and then using the `popkin` library in `R` (Ochoa and Storey 2021). We made no attempt to obtain a collection of kinship matrices each associated with "leaving a single chromosome out"; the computational details of the power simulations means that "leave one out" scans would have greatly increased the total computational burden.

Statistical tests for genotype/phenotype associations

We considered 2 tests for phenotype-genotype association: a marker- and a haplotype-based test. The marker-based test employs estimated dosages at single SNPs that take on values between 0 and 2 as predictors. In contrast, the haplotype-based test employs a vector of 8 dosages that sum to 2, with each element of the vector being the dosage of that particular haplotype (as returned from `Stitch`). A genome-wide marker-based scan examining 16,087,368 markers is time-consuming to carry out in the context of a power simulation. To carry out a genomewide scan without pruning SNPs, we implemented the "fast GWAS" approach of Ziyatdinov (Ziyatdinov *et al.* 2018) that is part of the `lme4qtl` package in `R` (https://github.com/variani/lme4qtl/blob/master/demo/gwas.R). The scan employs an EVD composition of the "null" model and kinship matrix to enable multiple SNPs to be tested for association in parallel, making the tests very fast. To avoid memory overflows we carried out the scan using the `blocksize` option, with resulting *P*-values transformed to $-\log_{10}$(*P*-values). Ziyatdinov's approach approximates fitting the following model with or without the effect of the SNP marker:

$$relmatLmer(formula = Y \sim (1 \mid ID) + SNP,\ relmat(ID = K),\ REML = TRUE).$$

(using `relmatLmer` from the `lme4qtl` package). We compared $-\log_{10}$(*P*-values) scores from the approximation relative to the full model and they were very similar.

We could not employ Ziyatdinov's approach for haplotype-based scans, as methods for speeding up GWAS scans assume biallelic single locus predictors, but haplotype tests have 8 predictors. We obtained a 10X speed-up by only carrying out a test at every tenth marker. The top 2 panels of Supplementary Fig. 1 depict haplotype scores over Chromosome 23 for 2 individuals. The figure shows that haplotype scores change slowly over megabase scales, and thus testing at every tenth marker is a reasonable compromise. We avoid fitting a full mixed model by creating a new phenotypic variable *Y'* as our dependent variable. *Y'* is derived by carrying out a principal component analysis on the kinship matrix, regressing the 15 largest principal component scores on *Y* (the actual phenotype), and retaining the residuals. Finally to avoid problems with co-linear predictors, we carried out a principal component analysis on the *NX8* matrix (*H*) of haplotype scores per locus, and retained only principal components scores accounting for greater than 5% of the total variation in the raw haplotype predictor matrix to create a new set of {independent} predictors (*H'*). We define *df* as the number of columns in *H'* ($df \leq 8$). We then obtain the residual sum of squares (*RSS*) from fits of 2 simple linear models: $Y' \sim H'$ and $Y' \sim 1$. An F-statistic is obtained by contrasting the *RSS* in the standard manner:

$$F = ((RSS_{null} - RSS_{full})/df)/(RSS_{full}/(N - df - 1)).$$

We finally derive a $-\log_{10}$(*P*-value) for each F. The computational advantage of this approach is that *Y'* need not be a vector of length *N*, but can instead be an *NXS* matrix consisting of *S* (where *S* is potentially >1,000) different simulated phenotypes. This strategy is efficient since much of the computational effort is inverting *H'*. This entire approach is meant to approximate fitting the full mixed model:

$$relmatLmer(formula = Y \sim (1 \mid ID) + H',\ relmat = list(ID = K),\ REML = FALSE)$$

using *stats::pchisq* and the number of columns in *H'* to compare the mixed model with and without the effect of *H'* to derive a $-\log_{10}$(*P*-value). Examining the $-\log_{10}$(*P*-value)'s from the full model compared to our approximation suggests that the approximation is valid.

Using false positives to set a significance threshold

For each % variation attributable to the QTL, genetic model, and statistical test we carried out 250 genome-wide scans (excluding Chromosome 19 markers), in which a causative "gene" was located on either Chromosome 23 or Chromosome 19. Thus, for any given candidate $-\log_{10}$(*P*-value) cutoff threshold, the number of hits associated with a causative locus on Chromosome 19 can be used to estimate the false-positive rate. We established thresholds of 6 and 7.5 respectively for the haplotype- and marker-based tests, these thresholds hold the genome-wide false positive rate at less than 5% (with some potential caveats discussed in results).

Power and localization

Given a threshold that controls the genome-wide false positive rate, we define a true positive hit as a replicate genome-wide scan in which at least 1 test statistic exceeds the significance threshold and is located on Chromosome 23. Power is then simply the percentage of true positive scans over the 250 replicate simulations carried out for any parameter combination. For each hit, we further calculate the distance between the most significant marker (MSM) and the causative SNP (or midpoint of the genetic interval defining a "gene") as a measure of localization.

Simulating pseudo-individuals

A reviewer of this manuscript provided code during the review process that allowed us to use the `*.EM.all.RData` files generated as part of the `Stitch`-based SNP imputation to simulate

pseudo-individuals consistent with observed genotypes and recombination rates. We implement this idea in the `sample_one_haplotype` function in the `sim.geno.R` script in the git. We used this approach to simulate 2,000 diploid Chromosome 23s. Haploid pseudo-individuals are created by recombining haploid alleles between *randomly chosen* individuals consistent with the rate at which those haplotypes break down in the observed dataset (and combining them at random to create diploids), thus any population structure (due to non random mating) present in the colony will tend to be eroded. Despite the potential shortcomings of this approach, allele frequencies and haplotype block lengths are effectively mimicked, so this strategy is an excellent starting point for simulating populations conditional on genotypes observed in an ascertained set of individuals. We suspect the resulting individuals are likely much more representative of an actual sample than would be obtained from a pure Monte Carlo approach. We used the simulated individuals to carry out Chromosome 23 only scans under all 3 genetic models with QTLs contributing 1%, 2.5%, or 5% to a complex trait. We do not simulate QTL on Chromosome 19, and instead employ the same $-\log_{10}$(*P*-value) thresholds as those used in the genome-wide scans described above. Focusing on a single chromosome allowed us to explore a range of population sizes via down-sampling (100, 250, 348, 500, 750, 1,000, 1,500, and 2,000) and estimate the impact of population size on power with a reasonable computational effort. Since we only explored Chromosome 23 and the pseudo-individuals likely do not model population structure present in the colony, we do not include background genetic variation when simulating phenotypes, nor do we generate or utilize kinship matrices in the model fitting.

Two demonstration traits

We measured "bleeding time" on 348 of the deermice employed in the above simulations. Each deermouse was also measured for several potential covariates, including: *Sex*, *Age* at time of assay, *Weight* at time of assay, and date of birth (*DOB*; in months from an arbitrary time in the past). Bleeding time is highly skewed with a long tail toward longer times, so the trait was quantile normalized to be normal (qn_Bleed). The model:

$$qn_{Bleed} \sim Sex + Age + Weight + DOB$$

showed Sex, Age, and DOB, but not Weight to be predictive of quantile normalized bleeding time, so we carry out all subsequent analyses on the residuals after Sex, Age, and DOB were removed. Visual examination of plots suggested that Age and DOB affect the trait as first order polynomials. Although bleeding time was measured in colony animals over 6 years, 75% of the animals were assayed over less than 2 years, but we do not know if our assay was changing slowly over time or the colony itself was changing. We similarly chose to examine the residuals obtained after modeling log transformed weight as a function of Sex, Age and a Sex by Age interaction. All 3 of these factors are strongly predictive of log(weight), although DOB is not.

We estimate the heritability of the residual quantile normalized bleeding phenotype or residual log(weight) and the earlier estimated kinship matrix using:

$$relmatLmer(formula = Y \sim (1 \mid IND),\ data = phen,$$
$$relmat = list(IND = K),\ REML = TRUE).$$

The heritability of the bleeding time phenotype was 6.5% and weight 62%. Although heritability estimates can be inaccurate given the $N = 348$ of this study, our estimates do suggest the heritability of bleeding time is low.

We carried out genome wide marker- and haplotype-based scans on quantile normalized bleeding time after Sex, Weight, and DOB were removed as well as residual log(Weight) using the significance threshold obtained from the simulations, created Manhattan plots and extracted the top 2 hits per character as evidenced from haplotype- and marker-based scores. We furthermore shuffled the phenotypes with respect to genotypes, carried out a second scan, created Manhattan plots, and created quantile-quantitative (QQ)-plots of the shuffled LOD scores vs actual LOD scores.

Platelet function candidate genes

We did an Ensembl Biomart query to extract the house mouse protein sequences associated with the 2 GO terms associated with platelet function (GO:0030168 and GO:0070527), retained the longest amino acid sequence per gene, blatted those proteins to the *P. leucopus* reference genome, and took the highest scoring hit per gene to obtain the locations of the 83 *P. leucopus* orthologous genes. We consider these platelet function genes to be candidate genes for a GWAS hit if they are within 2 Mb of the most-significant marker defining the top 4 hits. We additionally considered VKORC1 (@chr1:76732685-76735060) and its paralog VKORC1L1 (@chr23:21130849-21176651) as candidate genes. VKORC1 is associated with warfarin resistance in wild rodent populations. Given the 85 candidate genes, the loose 2 Mb criteria for a match, and 4 hits tested, we expect ~1 such hit by chance alone. Thus, the observation of a candidate gene within 2 Mb of a hit in and of itself is thus **not** strong evidence that that gene is causative. We did not attempt to identify an *a priori* set of candidate genes for log(weight), as the literature suggests several hundred such genes.

Results

Diploid genotypes can be accurately imputed in colony animals from low pass sequencing data

We carried out low coverage sequencing of 405 colony mice and aligned reads to the *P. leucopus* reference genome (Long *et al.* 2019). The upper panel of Fig. 1 is a histogram of raw sequence coverage obtained for the 348 individuals we study in depth in this work. Roughly 95% of the individuals have coverages between ~0.5X–2.5X; clearly such coverages are insufficient for directly calling diploid genotypes from raw reads. We used the program `Stitch` (Davies *et al.* 2016) to impute genotypes (and 8 pseudo-haplotypes) at 16,087,368 previously identified SNPs after filtering. To validate the genotype calls from `Stitch` we took advantage of 5 deermice for which we additionally obtained RNAseq data from the spleen. We aligned the RNAseq data to the genome, called genotypes, identified a set of 110,188 high-confidence genotypes at exonic SNPs, and compared these calls to dosage estimates at the same set of SNPs imputed from the low coverage data. The middle panel of Fig. 1 shows that only ~3–4% of SNPs are associated with an absolute error of greater than 0.10. The lower panel of Fig. 1 shows the value of the squared correlation coefficient (R^2) as a function of minor allele frequency. For SNPs with a minor allele frequency of greater than 5% imputation is excellent, even for the deermouse (ID=20694)

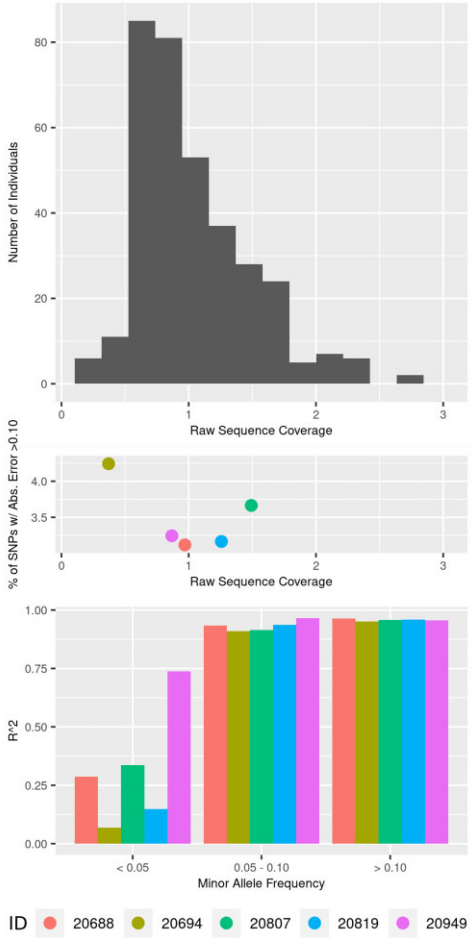

Fig. 1. Validation of genotype calls based on 5 individuals with high coverage RNAseq data. Top, histogram of raw sequence coverage for 348 individuals. Middle, % of SNPs with an absolute error >0.10 as a function of raw sequence coverage for the 5 validated samples. Bottom, R^2 as a function of minor allele frequency, colored by the 5 validation individuals.

sequenced to ~0.25X. For rarer alleles the quality of imputation is poorer and there is some evidence that the lowest coverage individual is associated with the poorest imputation. The performance of `Stitch` is largely consistent with the description of the algorithm (Davies *et al.* 2016), given the relatively modest sized panel we employ and $K = 8$ haplotypes. We attempted to compensate for our modest sample size by obtaining more coverage than recommended per individual, but accurate imputation of rarer alleles will generally require larger panels. Unlike a large study using `Stitch` to impute genotypes in outbred Swiss Webster (CFW) mice from Charles laboratory where $K = 4$ was sufficient for imputation and large genomic regions lack much variation (Nicod *et al.* 2016) (presumably due to a small number of founders and/or bottleneck), imputation in Peromyscus colony mice seems to benefit from a greater number of founder haplotypes. We explored different values of K, with $K = 8$ resulting in the best SNP imputation. Despite the colony being founded from 32 individuals, plots of founder haplotype frequencies among the 348 deermice examined here for Chromosome 23 (third panel of Supplementary Fig. 1) suggests much of that chromosome is segregating multiple common haplotypes. We further counted the number of haplotypes at a frequency of greater than 5% at every imputed position in the genome and observe the plurality of positions segregating 3 or 4 common variants, with roughly 5% segregating 6 or more founder alleles (bottom panel of Supplementary Fig. 1). It is perhaps counter-intuitive that high-quality diploid genotypes can be obtained routinely from animals sequenced to <1X, but imputation from low-pass sequencing data appears to be routine in *P. leucopus* closed colony deermice, as there is likely extensive linkage disequilibrium in the colony.

We estimated a kinship matrix from an evenly spaced subset of imputed and then hard-encored genome-wide SNPs (Supplementary Fig. 2). Although there is the potential for relatedness in the *P. leucopus* colony animals of our study, the kinship matrix of the subset 348 individuals we examine suggests that mating between closely related animals is somewhat avoided in the colony. Although some colony deermice appeared related to one another, and 2 known mother/pup pairs (indicated with the red arrow in Supplementary Fig. 2) were related at less than their expected 0.5. We did not attempt to remove these individuals from the study, nor did we rescale the kinship matrix to map QTL.

Simulations of the power to detect genes contributing to a complex trait in *P. leucopus* colony animals

To examine the power of different strategies to detect genes contributing to a complex trait in 348 colony animals, we simulated replicate scans using a set ~16M genome-wide (less Chromosome 19) imputed SNPs or 1/10th the number of imputed haplotypes to detect a causative QTL on either Chromosome 23 or Chromosome 19. Each replicate simulation defined a 100 kb "gene region" on either Chromosome 19 or 23 contributing to variation in a complex trait. For that region, we then considered 3 different genetic models per scan, which briefly included one with a single causative SNP, another with 10 causative SNPs in a 100 kb "gene," and a third in which all SNPs in a 100 kb "gene" are causative. Under all models, the causative gene contributed either 1%, 2.5%, or 5% to variation in a complex trait with trait a heritability of 50% and the remainder of the variation being modeled as polygenic genome-wide background genetic variation. A hit from a genome-wide scan where the causative gene is on Chromosome 19, but the scan lacks Chromosome 19 markers, allows us to estimate the false-positive rate. Whereas a hit from a genome-wide scan that localizes to Chromosome 23, where the causative QTL is also on Chromosome 23, allows us to estimate the true positive rate. Within each of the above combinatoric treatments (QTL on 19 vs 23; 3 genetic models; 3 QTL heritabilities) we simulated 250 replicate QTL locations. A genetic association with the complex trait was detected using a mixed model approach that leveraged the kinship matrix to control for population structure-based false positives, or a haplotype-based test that employed the set of 8 founder haplotypes generated by `Stitch` and controlled for relatedness by regressing out the first 15 principal components from the relatedness matrix on the simulated phenotype.

Figure 2 depicts QQ plots integrated over the 4,500 simulated genome-wide scans. Quantiles for QTL on Chromosome 19 represent a null distribution of $-\log_{10}$(*P*-values) each paired with a quantile for a QTL on Chromosome 23. In each panel, a departure from the line with a slope of 1 represents the "signal" in the experiment. It is apparent that there is considerably more signal as the % variation due to the causative QTL increases. In contrast, the particular genetic model simulated does not have a large impact on this power, which suggests genetic dissection of complex traits in this panel of deermice is somewhat robust to the genetic details of trait architecture. The observation that marker-based

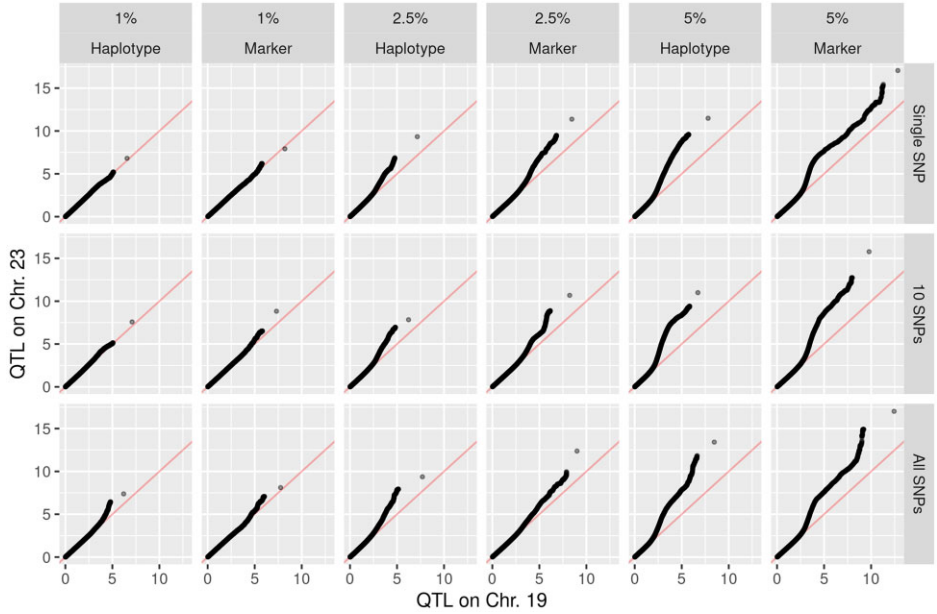

Fig. 2. QQ plots of $-\log_{10}$(*P*-values) over 250 genome-wide scans with a QTL located on either Chromosome 23 or 19 (control). Left to right considers 3 different contributions of the QTL to total variation, 1%, 2.5%, or 5%, and a haplotype- vs marker-based test; top to bottom considers 3 different genetic models: a single causative SNP, 10 causative SNPs in a 100 kb "gene," or all SNPs in a 100 kb gene. A line with a slope of 1 is plotted in red. The scan did not employ markers on Chromosome 19, so the observed departure from a straight line indicates a genetic signal.

Table 1. Properties of simulated genome scans to detect complex trait genes conditional on 348 genotyped deermice.

Percent variation	Causative Loci per gene	Test	FPR (%)[a]	TPR (%)[b]	Mean distance (kb)[c]
1.0	Single	SNP	0.8	0.8	11,750
1.0	Ten	SNP	0.0	1.2	959
1.0	All	SNP	1.2	0.8	640
2.5	Single	SNP	1.6	9.6	905
2.5	Ten	SNP	1.2	7.2	3,090
2.5	All	SNP	3.6	14.4	1,102
5.0	Single	SNP	9.2	50.8	918
5.0	Ten	SNP	6.0	48.0	1,118
5.0	All	SNP	8.4	50.0	969
		AVG	**3.6**		
1.0	Single	HAP[d]	2.0	0.0	NA
1.0	Ten	HAP	0.8	0.4	3,957
1.0	All	HAP	0.8	0.8	563
2.5	Single	HAP	2.0	3.2	2,383
2.5	Ten	HAP	2.0	3.2	4,161
2.5	All	HAP	1.6	7.6	2,305
5.0	Single	HAP	2.0	29.2	819
5.0	Ten	HAP	1.6	29.2	1,451
5.0	All	HAP	3.2	28.4	1,353
		AVG	**1.8**		

[a] False Positive Rate (FPR) is a Hit @ LOD $\geq$ 6|7.5 (QTL on Chr19).
[b] True Positive Rate (TPR) is a Hit @ LOD $\geq$ 6|7.5 (QTL on Chr23) and hit on Chr23.
[c] Mean distance between center of causative gene and MSM conditional on a hit.
[d] Haplotype-based.

tests generate larger $-\log_{10}$(*P*-values) than haplotype-based tests for QTL on either Chromosome 19 or 23 is not surprising. We carry out more marker- than haplotype-based tests and SNPs close to one another can often show less linkage disequilibrium than haplotype values.

Based on the QQ plots of Fig. 2, we defined $-\log_{10}$(*P*-value) thresholds for statistical significance for the marker- and haplotype-based tests as 7.5 and 6.0, respectively. Table 1 uses these thresholds to estimate both the true- and false-positive rates (FPR) as a function of the % variation due to the simulated QTL, the genetic model contributing to variation, or the test employed. For the marker-based tests the FPR appears constant among models if the % variation attributable to the QTL is held constant, but these data also suggest the FPR is an increasing function of the % variation at the causative SNP. The false-positive rate appears to be held below 5% per genome scan under a marker-based scan as long as there are not causative genes contributing $\geq$5% to complex trait variation (if such genes exist, they seem capable of generating false positive hits). In contrast, the threshold employed for the haplotype-based tests holds the false positive rate lower than 5% per genome scan irrespective of genetic model or the % variation due to the trait. We similarly estimated the true positive rate or power for a genome-wide scan as the proportion of scans with a causative gene on Chromosome 23

and a hit on the same chromosome. Irrespective of genetic model or statistical test the true positive rate is roughly equal to the false positive rate for QTL contributing 1% to trait variation. Power rises to perhaps 5–10% dependent on the statistical test for a QTL contributing 2.5% to trait variation and is considerably higher for a 5% QTL. To compare the statistical tests, while trying to account for the apparent elevated false-positive rate for the marker-based tests for large effect QTL, we calculate a false discovery rate averaged over genetic models as a function of statistical test for QTL contributing 2.5% or 5% to trait variation as approximately 25% and 10% respectively, with some indication false positive rates are lower for marker-based tests.

For simulation replicates with at least 1 marker exceeding the threshold and QTL explaining 2.5% or 5% of variation in a complex trait, we quantified localization ability. For a sample size of 348 deermice the distance between the causative site and the MSM is roughly 1–2.5 Mb. The average distance between the MSM and the causative site appears less for marker- than haplotype-based tests, and resolution seems to improve for QTL accounting for a greater fraction of phenotypic variation (Table 1). It also appears that the resolution is greater for genetic models in cases where a single locus is causative. The resolution does not appear to be a function of the observed $-\log_{10}$(*P*-value) for either statistical test over all genetic models within % variation attributed to the causative QTL (Supplementary Fig. 4). Localization ability in a GWAS on 348 colony deermice is at least an order of magnitude better than what could be achieved in a single generation QTL mapping experiment using a comparable number of animals (Lander and Botstein 1989).

Figure 3 depicts the 2 different statistical tests for 4 realizations of the simulation for causative "genes" on Chromosome 23 (and Supplementary Fig. 3 is a comparable plot for a control scan with the causative gene on Chromosome 19). Although very few replicate simulations exceeded the significance threshold for a QTL explaining 1% of phenotypic variation, the 2 such cases plotted seem to be detecting the simulated QTL. In the top panel despite the scan not being significant for the haplotype-based test there is a clear signal that just fails to reach the threshold centered on the location of the simulated QTL. We see similar patterns for 2 hits explaining 2.5% of phenotypic variation, with peaks in $-\log_{10}$(*P*-values) that seem to include the simulated causative variant. In all cases hits display 2–10 Mb "block-like" patterns of association, this is typical of many other examples (not shown). It is noteworthy, and perhaps counter-intuitive, that the marker-based test seems to be similarly powered to detect QTL whose underlying genetics is NOT a single causative site (Table 1). We suspect both tests are really detecting haplotypes or markers that happen to efficiently tag a large region, with the marker-based test not necessarily tagging the causative SNPs themselves. This is almost certainly true in the cases where the marker-based test has a hit under the all-SNPs model where the effects at individual SNPs are vanishingly small.

Since the examples of Fig. 3 were chosen from a much larger set of simulations to illustrate the power of this approach, we also attempted to show examples where the method fails or is misleading. Figure 4 depicts 2 examples of mis-mapped QTL in simulations where the QTL explains 5% of total variation. The top panel shows a QTL mapped to the left end of Chromosome 11, despite the simulated QTL being on Chromosome 23. This "hit" was significant using both the marker- and haplotype-based statistical test and exhibits the same block-like patterns of significance we expect of actual hits. One could not tell this is a false positive by simply examining Manhattan plots. The lower panel

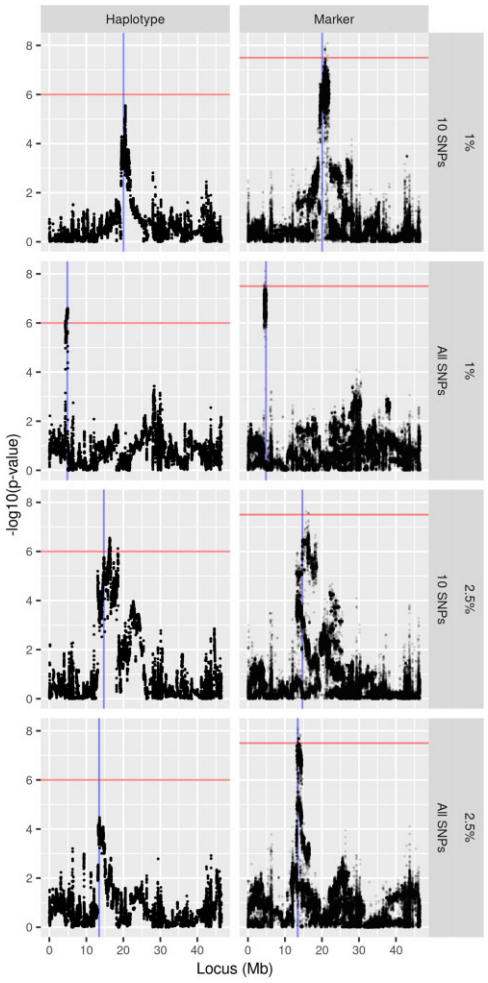

Fig. 3. Example of Chromosome 23 scans with a simulated QTL on Chromosome 23. Left, haplotype-based tests (red threshold = 6); right, marker-based tests (red threshold = 7.5); top to bottom are 2 different genetic models and 2 different contributions of the QTL contributing to total variation: 10 SNPs causative and all SNPs causative and 1% and 2.5%. The blue vertical line is the location of the causative gene.

depicts a similar situation where both tests detect a QTL toward the left end of Chromosome 23 that is not the location of the simulated QTL. It is interesting that in this case there is also considerable signal at the true location of the QTL, but the MSM marker is located ~20 Mb from the causative site. Figure 4 illustrates a curious and more general observation—significant false positives tend to occur near the ends of chromosomes. We do not have an explanation for why this occurs.

Simulating pseudo-individuals to examine the power and resolution of larger sized experiments

The simulated power of GWAS using 348 individuals is consistent with similar estimates associated with diversity outbred mice (Gatti *et al.* 2014), but is clearly modest. Despite many strengths, a weakness of power simulations conditional on a set of genotyped animals is that it is difficult to examine power at sample sizes greater than that of the ascertained population. We exploited haplotype and recombination rate information in `Stitch` output files to simulate 2,000 pseudo-individuals. The pseudo-individuals likely reproduce per region haplotype diversity and SNP site frequency spectrum, although they may not faithfully reproduce relatedness among individuals. Nonetheless, this

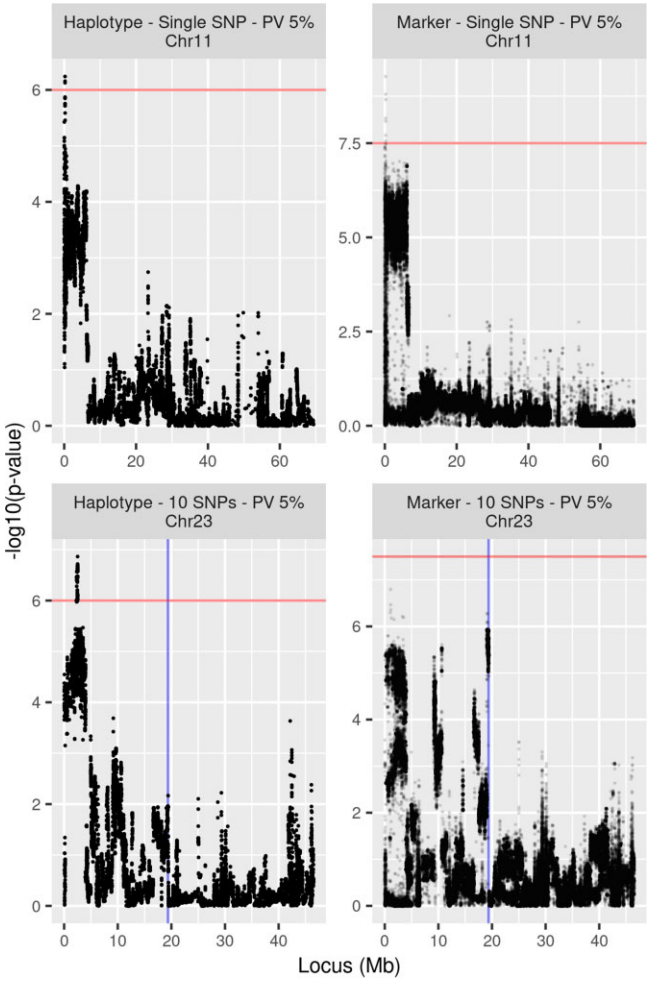

Fig. 4. Two scans demonstrating unusual results. Left, haplotype-based scans; right, marker-based scans. Top, the MSM is located on Chromosome 11, which is not the same chromosome as the causative gene (Chromosome 19), so this is a clear false positive hit. Bottom, the MSM is located at great distance (>16.8 Mb) from the causative gene location (indicated with a blue vertical line), that is the hit is on the correct chromosome but poorly localized.

method of simulating pseudo-individuals is promising enough that we carried out additional power simulations using these individuals using the same genetic models described above. To limit our computational burden, we only simulated QTL on Chromosome 23 and only carried out the scan using Chromosome 23 markers. Furthermore, we did not carry out the mixed model testing, nor did we simulate background genetic variation since the way individuals are created does not hold relatedness constant. Figure 5 depicts the power of GWAS experiments in our deermouse colony as a function of the underlying genetic model, variation contributed by the simulated QTL, and statistical test employed as a function of panel size using the significance threshold derived above. Power is largely determined by the number of individuals examined and the % variation attributable to the simulated QTL. For a QTL contributing 5% to variation in a complex trait power is quite good at $N = 1{,}000$ individuals, while for a more modest QTL contributing only 2.5% power approaches a reasonable 75% by 1,500 individuals and exceeds 90% by $N = 2{,}000$. These power profiles are consistent with mouse diversity outbred population simulations (Gatti *et al.* 2014). The power based on $N = 348$ pseudo-individuals appears to be slightly lower than those estimates obtained from simulations conditional on the actual 348 individuals genotyped under the haplotype-based test, and considerably lower than under the marker-based test, especially for QTL contributing 5% to variation in a complex trait.

If we condition on replicate simulations having significant hits the localization ability is potentially high in colony deermice. For a QTL contributing 2.5% to variation in a complex trait and $N = 1{,}000$, 1,500, or 2,000 individuals the average distance between the MSM and the causative site (or gene mid-point) was 200, 80, and 60 kb. These represent averages over the haplotype- and marker-based statistical tests and the all- and 10-SNP genetic models as they tended to perform similarly at a given sample size. Under the single causative SNP model and marker-based test the MSM was often the causative site, although we argue earlier that this situation is likely biologically not very plausible. For a QTL contributing 5% to variation in a complex trait the average distance between the MSM and the causative site was 50 kb averaged over the 2 multicausative site QTL models and appears independent of the number of animals examined. This suggests 50 kb is approximating the limits of resolution of experiments carried out in this colony.

Two demonstration traits

Three hundred and forty-eight of the genotyped animals were assayed for bleeding time and body weight. Bleeding time itself is not of great scientific interest, and does not appear to be highly heritable, but was easily measured on each animal at the same time a tail clip was carried out to obtain DNA for genotyping. Similarly, all animals were weighed before we obtained the tail clip. We transformed highly skewed raw bleeding times to be normally distributed and further removed the effects of sex, age (at time of assay), and date of birth from the trait before QTL mapping. We log transformed weight in grams and similarly removed the effects of sex, age (at time of assay), and their interaction. We carried out both a marker- and haplotype-based genome scan on the transformed traits, as well as a shuffled set of the same phenotypes, and employed thresholds for statistical significance derived above.

Supplementary Fig. 5 are QQ-plots of genome-wide $-\log10(P\text{-values})$ for bleeding time or log(weight) against permuted values of the same traits to mimic a null distribution. There appears to be only a small amount of signal for the bleeding time trait. Oddly log(weight) shows little signal under the marker-based test, but considerable signal under the haplotype-based test. The haplotype-based test further shows some evidence for genome-wide inflation for log(weight), but not for bleeding time, suggesting the first 15 principal components of the relatedness matrix are an effective control for structure for bleeding time, but not weight. Figure 6 shows Manhattan plots for the 2 demonstration traits under the 2 statistical tests. No hits exceed our significance thresholds apart from one region for log(weight) on Chromosome 3. It is of interest in the Manhattan plots that we see a few suggestive peaks, including some near the ends of chromosomes (which we believe are prone to false-positive hits). Finally, Supplementary Fig. 6 is a similar Manhattan plot of GWAS results, except carried out on permuted bleeding time or log(-weight), which serve as a negative control. These permuted scans generate several suggestive peaks (highlighting their danger), although no tests exceed the threshold for significance for either test and those suggestive peaks tend to be associated with chromosome ends.

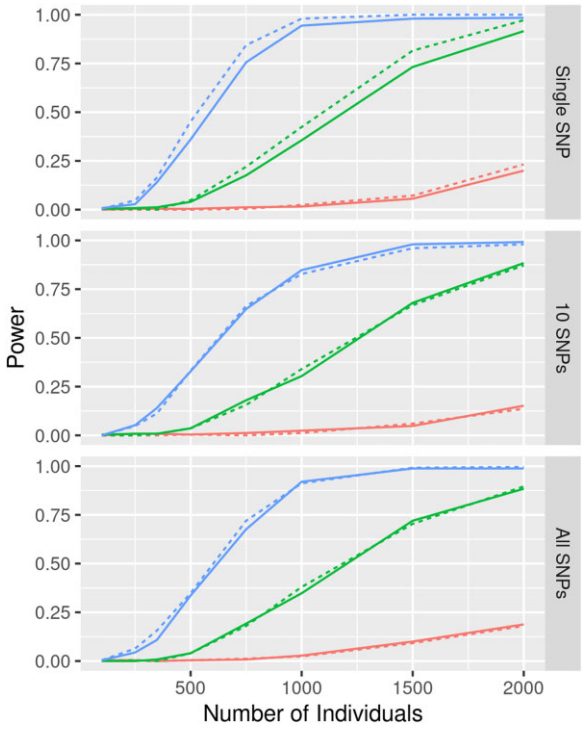

Fig. 5. Power of GWAS based on simulating 2000 pseudo-individuals down-sampled to examine smaller-sized experiments. The causative QTL is located on Chromosome 23 and the scan is carried out only for Chromosome 23. Dotted and solid lines are for marker- and haplotype-based tests using $-\log_{10}$(P-value) significance thresholds of 7.5 and 6.0, respectively. Blue, green, and red are for QTL accounting for 5%, 2.5%, and 1% of total variation, with all other variation being environmental. Panels from top to bottom are for causative genes with a single, 10, or all sites in a 100 kb region being causative.

The hits we identify in this study should only be considered suggestive [the haplotype-based test is significant for log(weight) but the statistic shows evidence for inflation]. Nonetheless, we more fully characterize the top 2 hits for each trait and test combination (Table 2). Figure 7 depicts Manhattan plots of a roughly 10 Mb region including the suggestive hit for each candidate, and Supplementary Fig. 7 are similar Manhattan plots except the entire chromosome harboring a hit is depicted. The haplotype-based test for log(weight) on Chromosome 17 and marker-based test for bleeding time on Chromosome 7 are uncharacteristic in displaying strong highly localized peaks without any flanking block-like signatures, and we suspect that, at least at this sample size, this methodology is perhaps prone to these artifacts. The other peaks are more typical and suggest some signal that is part of a larger block. The haplotype-based test hit on Chromosome 3 for log(weight) is interesting as there is an excellent candidate gene, *NPY*, only 2 Mb to its left at 67.9 Mb. Despite the strength of this candidate, the $-\log_{10}$(P-value) score falls off rapidly to the left of ~69.5 Mb, the score at the location of *NPY* is quite modest, and thus the details of linkage disequilibrium in this region seem to exclude this candidate. The marker-based hit for bleeding time on Chromosome 22 has a candidate gene (from an *a priori* set of 83 genes whose GO term matched platelet function), *Pip5K1c*, almost 5 Mb to the left of the MSM. Based on our simulations 5 Mb would be considered too far from the MSM to be considered a good candidate gene, except there is a secondary block of significance very close to *Pip5K1c*'s location (Fig. 7), so it is possible variation at *Pip5K1c* explains this peak. Given the modest support associated with these suggestive hits, carrying out a more highly powered experiment before any sort of functional characterization of these hits would seem prudent, and we view these candidate regions as illustrative of the approach and properties of colony deermice.

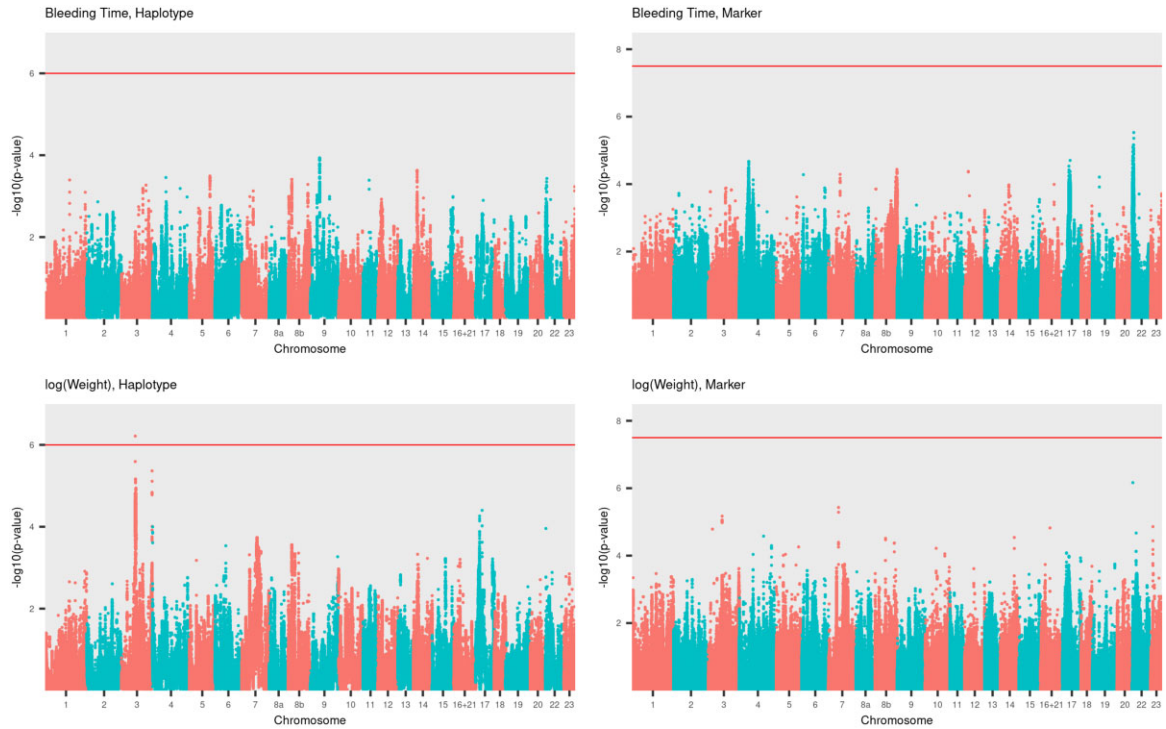

Fig. 6. Manhattan plots of genome-wide scans for the bleeding time and log(weight). Upper, bleeding time; lower, log(weight); left, haplotype-based scans; right, marker- based scans. Red line indicates the significance threshold.

Table 2. Genome scans to detect bleeding time or log(weight) causative regions.

		Discovery test[a]			Alternate test[b]	
Test	**Phenotype**	**Chr**	**MSM[c]**	**$-\log_{10}(p)$**	**MSM[c]**	**$-\log_{10}(p)$**
Haplotype	BT	9	43.348460	3.94	43.690184	2.98
Haplotype	BT	14	24.630059	3.63	29.097697	2.38
Haplotype	log(W)	3	69.841667	**6.21**	69.841637	5.17
Haplotype	log(W)	17	17.476222	4.40	17.478520	3.99
Marker	BT	4	49.414518	4.67	48.682012	2.27
Marker	BT	22	6.602023	5.53	6.477305	3.43
Marker	log(W)	7	45.898804	5.43	44.093750	2.11
Marker	log(W)	22	2.283682	6.17	2.283693	3.96

[a] Summary associated with test (column 1) used to detect the hit.
[b] Summary at the same location as test but for nondiscovery statistical test.
[c] MSM in Mb (digits after decimal define the bp).

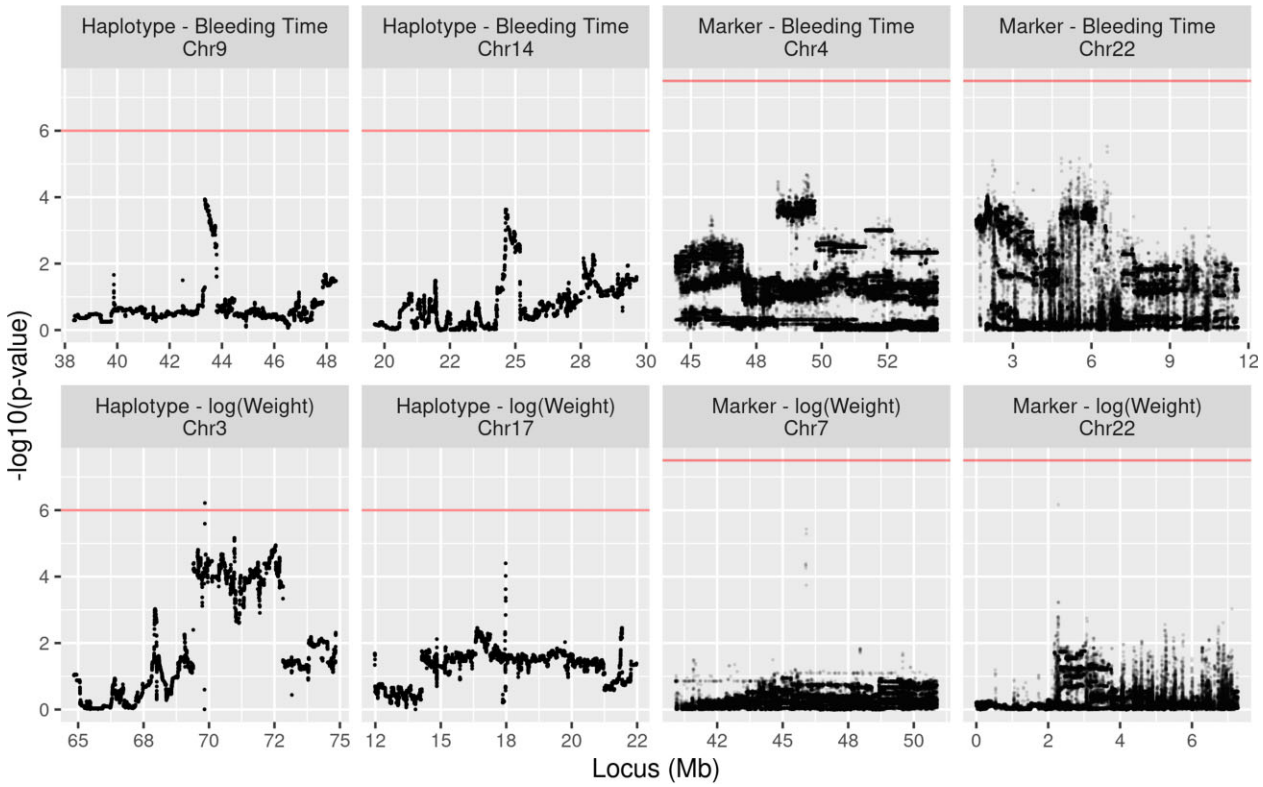

Fig. 7. Ten Mb regions centered around notable peaks from the genome-wide scans for the bleeding time and log(weight). Scans for bleeding time above and log(weight) below depict the regions with the 2 MSMs for either the haplotype- or marker-based test. A red line indicates the significance threshold.

Discussion

Here, we explore a diversity outbred approach for dissecting complex traits in *P. leucopus*, a nontraditional genetic model system. We take advantage of a colony of deer-mice founded from 38 individuals between 1982 and 1985 and subsequently maintained as a closed colony for 35+ years (~40–60 generations). We speculate that this *P. leucopus* colony shares many genetic features with DO/HS mice and rat populations. Most importantly, by virtue of its large number of founders we hypothesized the colony segregates many alleles per gene and given the many generations of colony maintenance initial linkage disequilibrium associated with the colony founding will be somewhat eroded. Unlike DO/HS mice and rats, where the strains used to found the panel are both known and highly characterized, the *P. leucopus* colony founders are both unknown and tissues were never archived. As a result, the DO mapping strategy of inferring known diploid founder haplotypes from marker data, and regressing on estimated haplotype dosages cannot be employed. Instead, phenotypes must be regressed onto a set of genome-wide imputed SNPs or local haplotypes inferred directly from the sample of genotyped individuals. Despite these potential shortcomings, we show that founder-unknown closed colonies harbor considerable mapping information that can be exploited to dissect complex traits.

We carried out low-pass sequencing on 405 mice (averaging ~1X per animal) and used `Stitch` (Davies *et al.* 2016) to impute diploid high-quality genotypes at ~16 million SNPs. We took advantage of 5 deer-mice also characterized via RNAseq to validate `Stitch`-based imputations and show that imputed genotypes are accurate. Our assumptions about the population structure of the deermouse colony were largely validated as much of the genome

appears to segregate 4 or more founder alleles. We carried out simulations to assess the ability to map QTL using deermice sampled from the colony under 3 different genetic models and for 3 different effect sizes at a causative QTL. The simulations made a single imputed SNP (or SNPs) in a 100 kb candidate gene region causative and then quantified the rate at which simulated genes could be detected via a genome-wide scan employing either a marker- or a novel haplotype-based statistical test. This is a classic simulation approach that has been widely used in several other contexts (c.f. McMullen *et al.* 2009; Aylor *et al.* 2011; King *et al.* 2012; Gatti *et al.* 2014). This simulation strategy has the attractive advantage of grounding the patterns of genetic variation, linkage disequilibrium, and relatedness between individuals. Conditional on the 348 deermice examined, the power to detect a QTL contributing to a complex trait was ~1–40%, largely dependent on the variation attributed to the simulated QTL, using a threshold that holds the false positive rate at less than 5% per genome scan. For the subset of simulation realizations with a significant hit, we were able to localize the causative site to ~2 Mb-sized windows. Although power and localization ability differed among the 2 tests utilized and the 3 genetic models, it did so only subtly. Despite the power of our approach being modest at 348 deermice, it appears to be comparable to those seen in diversity outbred mice and rats *conditional on sample size*. In fact, a study of power in DO mice populations shows almost identical power to the estimates of this paper (cf. Fig. 4 of Gatti *et al.* 2014), suggesting that like DO mice on the order of 500–1,000 mice are likely necessary to routinely map QTL contributing 5% to variation in the complex trait (Chitre *et al.* 2020).

A clear shortcoming of the simulation approach employed, is that it is not simple to simulate additional individuals from the colony to examine power and localization in a larger experiment. It is perhaps obvious that increasing the number of deermice examined will increase both power and localization ability. We used summary information generated as part of the `Stitch` imputation to generate 2,000 pseudo-individuals whose genotypes were consistent with the sample of deermice for which genotypes were obtained. Although the simulation approach does not perfectly model the colony, as it tends to create individuals equally related to one another, it likely does an excellent job of mimicking haplotype- and marker- frequency distributions and regional variation in these distributions. A pure Monte Carlo approach to simulating colony animals from first principles could be attempted, but simulations might not fully encapsulate variation in (unknown) founder haplotype frequencies nor longer range within chromosome patterns of LD, so would undoubtedly be a poorer representation. Using the 2,000 pseudo-individuals we carried out Chromosome 23 only scans to estimate the power of QTL mapping given increased sampling of deermice from the colony. There is considerable power to map QTL contributing 5% to variation in a complex trait given $N = 1{,}000$ deermice, or 2.5% to trait variation given $N = 1{,}500$–$2{,}000$ deermice. Furthermore, for experiments employing upwards of 1,000 deermice QTL were potentially localized to sub-megabase-sized regions. It is even conceivable the resolving ability of experiments carried out using *P. leucopus* colonies animals is higher than in more recently founded mouse or rat DO/HS populations.

We explored 3 genetic models and 2 statistical tests. Although one of our models simulates the customary single SNP contributing ~5% to variation in a complex trait, in light of human GWAS studies generally failing to identify intermediate frequency alleles of large effect (Manolio *et al.* 2009), this model is almost certainly wrong much of the time. We thus chose to simulate 2 additional models in which variation at a gene contributing to a complex trait was due to 10 or potentially hundreds of causative SNPs in that gene. Such models are broadly consistent with the idea of mutation selection balance maintaining variation at complex trait genes (Pritchard 2001; Thornton *et al.* 2013). Although our simulations are undoubtedly an over-simplification of such multiple causative site models, they may be useful for comparing marker- vs haplotype-based tests. We hypothesized that single marker tests would perform well under the single causative site model (but not necessarily multisite models), whereas haplotype-based tests would perform well under multiple causative site models. Instead, we observe that both marker- and haplotype-based statistical tests performed similarly irrespective of the underlying genetic model. The megabase-sized blocks of elevated marker-based LOD scores associated with the location of simulated QTL suggests that LD is extensive enough in colony mice that any given causative region is tagged by several hundred SNPs.

We finally carried out a GWAS on 2 demonstration traits, bleeding time following a small tail clip and log-transformed body weight. We detected a single significant region under the haplotype-based test and log(weight). Despite being significant, we only consider this hit as suggestive, as the log(weight) trait for the haplotype-based test showed some inflation of $-\log_{10}$(*P*-value) test statistics. We more carefully examined the 2 highest scoring regions for each statistical test and trait and identified a candidate gene, *Pip5K1c*, consistent with the $-\log_{10}$(*P*-value) test statistic for the marker-based test on Chromosome 22 for bleeding time and the only *a priori* candidate gene co-localizing with our suggestive hits.

Although *Peromyscus* is not considered a genetic model system, the genus is well suited for addressing several questions of biologist interest including the genetic bases of longevity, behavior, physiology, adaptation, and a deermouse's ability to serve as a disease vector. Each of these phenotypes are associated with clear scientific and/or human disease questions where the ability to dissect complex traits in *Peromyscus* would be of great value. Our results suggest that individuals from a long-maintained colony of *P. leucopus* deermice can be utilized in a manner very similar to DO/HS mice and rats for dissecting complex traits. Although >1,000 deermice should be measured to consistently dissect typical complex traits.

The results of this study are not only of interest to *Peromyscus* researchers. The simulations of this work show that virtually any closed colony of animals derived from a limited number of founders and maintained for several dozen generations can be used to dissect complex traits using a DO/HS-style approach. This can be accomplished cost effectively using low-pass short-read sequencing and genotype imputation, as suggested by Nicod *et al.* (2016), provided the species exhibits high levels of nucleotide diversity and has a reference genome. Colonies derived from unknown founders clearly have some disadvantages but having unknown-founders is not a complete impediment. Given an annotated genome and a GWAS scan, candidate genes can be identified, and the genetic structure of the colony characterized with respect to founder haplotypes for small regions of interest. It is also possible sophisticated extensions to `Stitch` (Davies *et al.* 2016) will eventually allow for phased haplotype estimates from large samples from closed colonies. Our work should further motivate a founder unknown DO/HS-strategy for other species for which closed colonies exist.

Data availability

Raw sequencing reads have been uploaded to the SRA under project PRJNA751054 and accessions SAMN20504476 through SAMN20504850 plus a blank (SAMN20504851). We host the filtered genotype calls (SNP and haplotype calls, and IDs) at: http://wfitch.bio.uci.edu/tdlong/sandvox/publications.html. There is a github archive with code to reproduce analyses: https://github.com/tdlong/Pero_power_revise.git.

Supplemental material is available at *GENETICS* online.

Funding

Work was funded by National Institutes of Health (NIH) grants R21AI126037 and R01AI57513 to ADL and AGB. This work was made possible, in part, through access to the Genomics High Throughput Facility Shared Resource of the Cancer Center Support Grant (P30CA-062203) at the University of California, Irvine and NIH shared instrumentation grants 1S10RR025496-01, 1S10OD010794-01, and 1S10OD021718-01.

Conflicts of interest

None declared.

Literature cited

Aylor DL, Valdar W, Foulds-Mathes W, Buus RJ, Verdugo RA, Baric RS, Ferris MT, Frelinger JA, Heise M, Frieman MB, *et al.* Genetic analysis of complex traits in the emerging Collaborative Cross. Genome Res. 2011;21(8):1213–1222. doi:10.1101/gr.111310.110.

Barbour AG. Infection resistance and tolerance in *Peromyscus* spp., natural reservoirs of microbes that are virulent for humans. Semin Cell Dev Biol. 2017;61:115–122. doi:10.1016/j.semcdb.2016.07.002.

Barbour AG, Shao H, Cook VJ, Baldwin-Brown J, Tsao JI, Long AD. Genomes, expression profiles, and diversity of mitochondria of the White-footed Deermouse *Peromyscus leucopus*, reservoir of Lyme disease and other zoonoses. Sci Rep. 2019;9(1):17618.

Bedford NL, Hoekstra HE. Peromyscus mice as a model for studying natural variation. Elife. 2015;4:e06813. doi:10.7554/eLife.06813.

Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169(7):1177–1186. doi:10.1016/j.cell.2017.05.038.

Broze GJ Jr, Yin ZF, Lasky N. A tail vein bleeding time model and delayed bleeding in hemophiliac mice. Thromb Haemost. 2001;85(4):747–748. doi:10.1055/s-0037–1615666.

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005–D1012. doi:10.1093/nar/gky1120.

Chitre AS, Polesskaya O, Holl K, Gao J, Cheng R, Bimschleger H, Garcia Martinez A, George T, Gileta AF, Han W, *et al.* Genome-wide association study in 3,173 outbred rats identifies multiple loci for body weight, adiposity, and fasting glucose. Obesity (Silver Spring)). 2020;28(10):1964–1973. doi:10.1002/oby.22927.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al.*; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–2158. doi:10.1093/bioinformatics/btr330.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, *et al.* Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008. doi:10.1093/gigascience/giab008.

Davies RW, Flint J, Myers S, Mott R. Rapid genotype imputation from sequence without reference panels. Nat Genet. 2016;48(8):965–969. doi:10.1038/ng.3594.

de Koning D-J, McIntyre LM. Back to the future: multiparent populations provide the key to unlocking the genetic basis of complex traits. Genetics. 2017;206(2):527–529. doi:10.1534/genetics.117.203265.

Gatti DM, Svenson KL, Shabalin A, Wu L-Y, Valdar W, Simecek P, Goodwin N, Cheng R, Pomp D, Palmer A, *et al.* Quantitative trait locus mapping methods for diversity outbred mice. G3 (Bethesda). 2014;4:1623–1633. doi:10.1534/g3.114.013748.

Hansen C, Spuhler K. Development of the National Institutes of Health genetically heterogeneous rat stock. Alcohol Clin Exp Res. 1984;8(5):477–479. doi:10.1111/j.1530-0277.1984.tb05706.x.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–915. doi:10.1038/s41587-019–0201-4.

King EG, Macdonald SJ, Long AD. Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. Genetics. 2012;191(3):935–949. doi:10.1534/genetics.112.138537.

Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. 1989;121(1):185–199. doi:10.1093/genetics/121.1.185.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN]. 2013.

Long AD, Baldwin-Brown J, Tao Y, Cook VJ, Balderrama-Gutierrez G, Corbett-Detig R, Mortazavi A, Barbour AG. The genome of *Peromyscus leucopus*, natural host for Lyme disease and other emerging infections. Sci Adv 2019;5(7):eaaw6441.

Long AD, Macdonald SJ, King EG. Dissecting complex traits using the Drosophila synthetic population resource. Trends Genet. 2014;30(11):488–495. doi:10.1016/j.tig.2014.07.009.

Macdonald SJ, Long AD. Joint estimates of quantitative trait locus effect and frequency using synthetic recombinant populations of *Drosophila melanogaster*. Genetics. 2007;176(2):1261–1281. doi:10.1534/genetics.106.069641.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, *et al.* Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–753. doi: 10.1038/nature08494.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–1303. doi:10.1101/gr.107524.110.

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, *et al.* Genetic properties of the maize nested association mapping population. Science. 2009;325(5941):737–740. doi: 10.1126/science.1174320.

Milovic A, Bassam K, Shao H, Chatzistamou I, Tufts DM, Diuk-Wasser M, Barbour AG. *Lactobacilli* and other gastrointestinal microbiota of *Peromyscus leucopus*, reservoir host for agents of Lyme disease and other zoonoses in North America. PLoS One. 2020;15(8):e0231801. doi:10.1371/journal.pone.0231801.

Mott R, Talbot CJ, Turri MG, Collins AC, Flint J. A method for fine mapping quantitative trait loci in outbred animal stocks. Proc

Natl Acad Sci USA. 2000;97(23):12649–12654. doi: 10.1073/pnas.230304397.

Najjar DA, Normandin AM, Strait EA, Esvelt KM. Driving towards ecotechnologies. Pathog Glob Health. 2017;111(8):448–458. doi: 10.1080/20477724.2018.1452844.

Nicod J, Davies RW, Cai N, Hassett C, Goodstadt L, Cosgrove C, Yee BK, Lionikaite V, McIntyre RE, Remme CA, *et al.* Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. Nat Genet. 2016;48(8):912–918. doi: 10.1038/ng.3595.

Ochoa A, Storey JD. Estimating FST and kinship for arbitrary population structures. PLoS Genet. 2021;17(1):e1009241. doi: 10.1371/journal.pgen.1009241.

Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. 2001;69(1):124–137. doi: 10.1086/321272.

Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet. 2002;11(20):2417–2423. doi:10.1093/hmg/11.20.2417.

Svenson KL, Gatti DM, Valdar W, Welsh CE, Cheng R, Chesler EJ, Palmer AA, McMillan L, Churchill GA. High-resolution genetic mapping using the mouse diversity outbred population. Genetics. 2012;190(2):437–447. doi:10.1534/genetics.111.132597.

The Peromyscus Genetics Stock Center. Peromyscus leucopus White-footed mouse LL Stock. The Peromyscus Genetics Stock Center. 2021. https://sc.edu/study/colleges_schools/pharmacy/centers/peromyscus_genetic_stock_center.

Thornton KR, Foran AJ, Long AD. Properties and modeling of GWAS when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. PLoS Genet. 2013;9(2): e1003258. doi:10.1371/journal.pgen.1003258.

Woods LCS, Mott R. Heterogeneous stock populations for analysis of complex traits. Methods Mol Biol. 2017;1488:31–44.

Ziyatdinov A, Vázquez-Santiago M, Brunel H, Martinez-Perez A, Aschard H, Soria JM. lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. BMC Bioinformatics. 2018;19(1):68. doi:10.1186/s12859-018–2057-x.

Communicating editor: J. Flint