A selfish genetic element drives recurring selective sweeps in the house mouse

John P Didion[1,2,3*], Andrew P Morgan[1,2,3*], Liran Yadgary[1,2,3], Timothy A Bell[1,2,3], Rachel C McMullan[1,2,3], Lydia Ortiz de Solorzano[1,2,3], Janice Britton-Davidian[4], Carol J Bult[5], Karl J Campbell[6,7], Riccardo Castiglia[8], Yung-Hao Ching[9], Amanda J Chunco[10], James J Crowley[1], Elissa J Chesler[5], John E French[11], Sofia I Gabriel[12], Daniel M Gatti[5], Theodore Garland Jr.[13], Eva B Giagia-Athanasopoulou[14], Mabel D Giménez[15], Sofia A Grize[16], İslam Gündüz[17], Andrew Holmes[18], Heidi C Hauffe[19], Jeremy S Herman[20], James M Holt[21], Kunji Hua[1], Wesley J Jolley[22], Anna K Lindholm[16], María J López-Fuster[23], George Mitsainas[14], Maria Mathias[23], Leonard McMillan[21], M Graça Ramalhinho[23], Barbara Rehermann[24], Stephan P Rosshart[24], Jeremy B Searle[12], Meng-Shin Shiao[25], Emanuela Solano[8], Karen L Svenson[5], Pat Thomas-Laemont[10], David W Threadgill[26], Jacint Ventura Queija[27], George M Weinstock[28], Daniel Pomp[1,3], Gary A Churchill[5], Fernando Pardo-Manuel de Villena[1,2,3]

1. Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, US

2. Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, US

3. Carolina Center for Genome Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, US

4. Institut des Sciences de l'Evolution, Université de Montpellier, CNRS, IRD, EPHE, Montpellier, FR

5. The Jackson Laboratory, Bar Harbor, ME, US

6. Island Conservation, Puerto Ayora, Galápagos Island, EC

7. School of Geography, Planning & Environmental Management, The University of Queensland, St Lucia, AU

8. Department of Biology and Biotechnologies "Charles Darwin", University of Rome "La Sapienza", Rome, IT

9. Institute of Zoology, National Taiwan University, Taipei, TW

10. Department of Environmental Studies, Elon University, Elon, NC, US

11. National Toxicology Program, National Institute of Environmental Sciences, NIH, Research Triangle Park, NC, US

12. Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, US

13. Department of Biology, University of California Riverside, Riverside, CA, US

14. Section of Animal Biology, Department of Biology, University of Patras, Patras, GR

15. Instituto de Biolog'a Subtropical, CONICET Ð Universidad Nacional de Misiones, Posadas, MS, AR

16. Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, CH

17. Department of Biology, Faculty of Arts and Sciences, University of Ondokuz Mayis, Samsun, TU

18. Laboratory of Behavioral and Genomic Neuroscience, National Institute on Alcohol Abuse and Alcoholism, NIH, Bethesda, MD, US

19. Department of Biodiversity and Molecular Ecology, Centre for Research and Innovation, Fondazione Edmund Mach, S. Michele all'Adige, TN, IT

20. Department of Natural Sciences, National Museums Scotland, Edinburgh, UK

21. Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, US

22. Island Conservation, Santa Cruz, CA, US

23. Department of Animal Biology & Centre for Environmental and Marine Studies, Faculty of Science, University of Lisbon, Lisboa, PT

24. Immunology Section, Liver Diseases Branch, National Institute of Diabetes and Digestive and Kidney Diseases, NIH, Bethesda, MD, US

25. Research Center, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, 10400, Thailand , Bangkok, TH

26. Department of Veterinary Pathobiology and Department of Molecular and Cellular Medicine, Texas A&M University, College Station, TX, US

27. Departament de Biologia Animal, Biologia Vegetal i Ecologia, Facultat de Ciencies, Universitat Autonoma de Barcelona, Barcelona, ES

28. Jackson Laboratory for Genomic Medicine, Farmington, CT, US

1 * These authors contributed equally to this work

2

Introduction (264 words)

A selective sweep is the result of strong positive selection rapidly driving newly occurring or standing genetic variants to fixation, and can dramatically alter the pattern and distribution of allelic diversity in a population or species. Population-level sequencing data have enabled discoveries of selective sweeps associated with genes involved in recent adaptations in many species[1-6]. In contrast, much debate but little empirical evidence addresses whether "selfish" genes are capable of fixation – thereby leaving signatures identical to classical selective sweeps – despite being neutral or deleterious to organismal fitness[7-11]. Here we show that *R2d2*, a large copy-number variant that causes non-random segregation of mouse Chromosome 2 in females due to meiotic drive[12], has driven recurring selective sweeps while having no discernable effect on fitness. We tested multiple closed breeding populations from six outbred backgrounds and found that alleles of *R2d2* with high copy number ($R2d2^{HC}$) rapidly increase in frequency, and in most cases become fixed in significantly fewer generations than can be explained by genetic drift. A survey of 16 natural mouse populations in Europe and the United States revealed that $R2d2^{HC}$ alleles are circulating at intermediate frequencies in the wild; moreover, patterns of local haplotype diversity are consistent with recent positive selection. Our data provide direct evidence of populations actively undergoing selective sweeps driven by a selfish genetic element, and demonstrate that meiotic drive can rapidly alter the genomic landscape in favor of mutations with neutral or even negative effect on overall Darwinian fitness. Further study and updated models are required to

clarify the relative contributions of selfish genes, adaptation and genetic drift to evolution.

Main text (1710 words)

With few exceptions[13,14], evolution is viewed through the lens of history, by inference from the comparison of genetically distinct populations that are thought to share a common origin. Much evidence suggests that novel or standing genetic variants can be rapidly fixed by strong positive selection if they are beneficial to organismal fitness. A classic (or "hard") selective sweep describes the process of a newly arising mutation with large positive fitness effect increasing in frequency in a population, ultimately leading to the fixation of the mutation. The concept was later expanded to include "soft" selective sweeps in which selection acts on standing variation in the advent of a change in environment[15,16]. As a selected variant rises in frequency, it carries with it linked genetic variation ("genetic hitchhiking"), thereby reducing local haplotype diversity. This signature – reduced genetic diversity relative to the neutral expectation in a region of linkage disequilibrium (LD) surrounding an advantageous allele – allows retrospective identification of selective sweeps in samples of contemporaneous populations.

In most reported selective sweeps, candidate regions contain genes (or sets of related genes) whose roles in organismal fitness are obvious. Prominent examples include alleles at the *Vkorc1* locus, which confers rodenticide resistance in the brown rat[17], and enhancer polymorphisms conferring lactase persistence in human beings[1]. However, a selective sweep may also be driven

by a “selfish” allele that is only beneficial to itself[18], as has been suggested with *Segregation Distorter* in *Drosophila*[3] and transmission distortion in domestic chickens[19].

We previously reported a novel meiotic drive responder locus (*R2d2*) whose core is a variably sized copy number gain on mouse Chromosome 2 that contains a single annotated gene (*Cwc22*, a spliceosomal protein). Females that are heterozygous at *R2d2* preferentially transmit to their offspring the allele with high copy number ($R2d2^{HC}$) relative to the allele with low copy number ($R2d2^{LC}$) to an extent that depends on genetic background. Distorted transmission of $R2d2^{HC}$ is also either uncorrelated or negatively correlated with fecundity – a major component of absolute fitness – depending on genetic background[12]. $R2d2^{HC}$ therefore behaves as a selfish genetic element. Here, we tested the hypothesis that this element is capable of causing selective sweeps in both laboratory and wild populations of house mice.

The Diversity Outbred (DO) is a randomized outbreeding population derived from eight inbred mouse strains that is maintained under conditions designed to minimize the effects of both selection and genetic drift. Expected time to fixation or loss of an allele present in the founder generation (with initial frequency 1/8) is ~900 generations[20]. The WSB/EiJ founder strain contributed an $R2d2^{HC}$ allele which underwent a more than three-fold increase (from 0.18 to 0.62) in 13 generations ($p < 0.001$ by simulation; range 0.03 – 0.26 after 13 generations in 1000 simulation runs) (**Figure 1A**), accompanied by distorted allele frequencies across a ~100 Mb region linked to the allele (**Figure 1B**). Litter sizes in the DO

were approximately constant during the increase in *R2d2*^{HC} frequency (mean 7.48 ± 0.27; **Figure 1A**), suggesting that *R2d2* does not impact overall reproductive fitness in this population.

We also observed selective sweeps in selection lines derived from the ICR:Hsd outbred population[21], in which *R2d2*^{HC} alleles are segregating (**Figure 1C**). Three of four lines selectively bred for high voluntary wheel-running (HR lines) and two of four control lines (10 breeding pairs per line per generation in both conditions) went from starting *R2d2*^{HC} frequencies ~0.75 to fixation in 60 generations or less: two lines were fixed by generation 20, and three more by generation 60. In simulations mimicking this breeding design and neutrality (**Extended Data Fig. 1**), median time to fixation was 46 generations (5th percentile: 9 generations). Although the *R2d2*^{HC} allele would be expected to eventually fix by drift in 6 of 8 lines given its high starting frequency, fixation in two lines within 20 generations and three more lines by 60 generations is not expected (p = 0.003 by simulation). In a related advanced intercross segregating for high and low copy number alleles at *R2d2* (HR8xC57BL/6J[22]), we observed that *R2d2*^{HC} increased from a frequency of 0.5 to 0.85 in just 10 generations and fixed by 15 generations, versus a median 184 generations in simulations (p < 0.001) (**Figure 1D**). The increase in *R2d2*^{HC} allele frequency in the DO and the advanced intercross populations occurred at least an order of magnitude faster than what is predicted by drift alone.

Using archival tissue samples, we were able to determine *R2d2* allele frequencies in the original founder populations of 6 of the ~60 wild-derived

laboratory strains in common use[23]. In four strains, WSB/EiJ, WSA/EiJ, ZALENDE/EiJ, and SPRET/EiJ, *R2d2*^{HC} alleles were segregating in the founders and are now fixed in the inbred populations. In the other two strains, LEWES/EiJ and TIRANO/EiJ, the founders were not segregating for *R2d2* copy number and the inbred populations are fixed for *R2d2*^{LC} (**Extended Data Fig. 2**). This trend in wild-derived strains is additional evidence of the tendency for *R2d2*^{HC} to go to fixation in closed breeding populations when segregating in the founder individuals.

Recently, whole-genome sequencing revealed extreme copy number variation at the *R2d2* locus in a sample of eight mice trapped in the Cologne-Bonn region of Germany[24]. To determine more broadly the distribution and frequency of *R2d2* alleles in wild mice, we assayed *R2d2* copy number in 396 individuals sampled from 14 European countries and the United States (JPD, JBS, and FPMV, in preparation) (**Supplementary Table 1** and **Extended Data Fig. 3A**). We found that *R2d2*^{HC} alleles are segregating at a wide range of frequencies in nature (0.00 – 0.67; **Supplementary Table 2**).

To examine patterns of haplotype diversity around *R2d2*, we genotyped the wild-caught mice at 77,808 SNPs on the medium-density MegaMUGA array[25,26]. Conventional tests[27,28] failed to detect a selective sweep around *R2d2* (**Extended Data Fig. 4**). However, the power of these tests is limited when the favored allele is common in the ancestral population, when a sweep is ongoing, or when linkage disequilibrium is weak[29]. In the case of very recent or strong positive selection, unrelated individuals are more likely to share extended

segments identical by descent (IBD) in the vicinity of the selected locus[30], compared with a population subject only to genetic drift. Consistent with this prediction, we observed a significant excess of shared IBD across populations around *R2d2* (**Figure 2A**): *R2d2* falls in the top 0.25% of IBD-sharing scores across the autosomes. In all cases, the shared haplotype has high copy number. Strong signatures are also evident at a previously identified target of positive selection, the *Vkorc1* locus (distal Chromosome 7)[31].

In principle, the strength and age of a selective sweep can be estimated from the rate of LD decay around the locus under selection. From the SNP data, we identified a ~1 Mb haplotype with significantly greater identity between individuals with $R2d2^{HC}$ alleles compared to the surrounding sequence. We used published sequencing data from 26 wild mice[24] to measure LD decay around *R2d2* and found that the haplotypes associated with $R2d2^{HC}$ alleles are longer than those associated with $R2d2^{LC}$ (**Figure 2B-C**). This pattern of haplotype homozygosity is consistent with positive selection over an evolutionary timescale as short as 450 generations. However, we note that $R2d2^{HC}$ alleles are refractory to recombination in laboratory crosses[12], and a nearly identical 2 – 5 Mb haplotype (0.5 – 1.1 cM in the standard mouse genetic map) is shared by several classical and wild-derived inbred strains that have different karyotypes and whose ancestors are separated by at least 10,000 generations[32] (**Extended Data Fig. 5**).

The discrepancy between the degree of transmission distortion in favor of $R2d2^{HC}$ in laboratory populations (up to 95%) and its moderate allele frequency

in the wild (0.14 worldwide) is initially surprising. However, in contrast to most other known meiotic drive systems, in which the component elements are tightly linked, the action of *R2d2*^{HC} is dependent on genetic background at multiple unlinked "modifier" loci[12]. Since the identities of these modifiers are currently unknown, we cannot predict their frequencies or distributions in the wild; thus, there is no reason to expect *R2d2* to be monomorphic. We used forward-in-time simulations to explore the population dynamics of meiotic drive in the simple case of two unlinked modifier loci. Assuming an additive model, we found that fixation of a focal allele (*e.g. R2d2*^{HC}) by meiotic drive was no more frequent than under the null model of neutral drift when permissive modifier alleles were rare, except when effective population size was large. An epistatic model required even greater modifier allele frequencies and/or population sizes to fix a focal allele (**Extended Data Fig. 6**). The maintenance of closely related *R2d2*^{HC} haplotypes at intermediate frequencies in multiple temporally and spatially diverged subpopulations (as we observed in mice of both European and American origin) is consistent with a model in which the stochastic and unlinked fluctuation of the *R2d2* and modifier alleles, along with the overdominant nature of meiotic drive[33], establish the conditions necessary for balancing selection[34].

Although a selfish selective sweep has clear implications for such experimental populations as the DO and the Collaborative Cross[12], the larger evolutionary implications of selfish sweeps are less obvious. On one hand, selective sweeps may be relatively rare, as appears to be the case for classic selective sweeps in recent human history[35]. On the other hand, theory and comparative studies

indicate that centromeric variants can act as selfish elements subject to meiotic drive[9,36] and be a potent force during speciation[8,18,33]. The fate of a selective sweep due to a selfish element depends on the fitness costs associated with the different genotypic classes. For example, maintenance of intermediate frequencies of the *t*-complex[37] and *Segregation Distorter*[38] chromosomes in natural populations of mice and *Drosophila*, respectively, is thought to result from decreased fecundity associated with those selfish elements. Further study will be required to elucidate the fitness effects of $R2d2^{HC}$ and its associated haplotype in the wild.

Evolutionary dogma holds that a newly arising mutation's likelihood of becoming established, increasing in frequency and even going to fixation within a population is positively correlated with its effect on organismal fitness. Here, we have provided evidence of a selfish genetic element driving recurring selective sweeps in which change in allele frequency and effect on organismal fitness are decoupled. This has broad implications for evolutionary studies: independent evidence is required to determine whether loci implicated as drivers of selective sweeps are adaptive or selfish.

Online Methods

Mice

Diversity Outbred (DO): All DO mice are bred at The Jackson Laboratory in waves (or "generations") lasting ~3 months. Some offspring from each generation are used as founders for subsequent generations. Pedigrees are used to identify mating pairs that minimize the chances for natural selection to occur. Individual investigators purchased mice (**Supplementary Table 3**) for unrelated studies, and contributed either tissue samples or genotype data to this study. All mice were handled in accordance with the IACUC protocols of the investigators' respective institutions.

High running (HR) selection lines: The breeding and selection scheme of the HR lines is described elsewhere[21]. Briefly, two generations prior to selection (generation -2), offspring of a base population of ICR:Hsd outbred mice were randomly assigned to 112 mating pairs. The offspring of those pairs were used as founders for eight lines (10 breeding pairs per line). At each generation thereafter, within-family selection for voluntary wheel running was performed: the highest-running male and female from each family were randomly paired (avoiding sibling matings) to produce the next generation.

HR8xC57BL/6J advanced intercross: The production of the HR8xC57BL/6J advanced intercross is described elsewhere[39,40]. Briefly, at ~8 wk of age, progenitor HR8 mice (HR line #8, 44th generation of artificial selection for high voluntary wheel running) and C57BL/6J (B6) mice underwent a reciprocal cross breeding protocol. 22 males and 22 females per line produced the F1 generation,

and three subsequent generations (F2, G3, G4) were derived from the two reciprocal mating types (B6 males × HR8 females and B6 females × HR8 males). Once established, the two reciprocal cross-line populations were not mixed. In total, 32 mating pairs from each reciprocal cross population were established each generation. To avoid inbreeding and increase the effective population size, interfamilial matings were assigned each generation utilizing a Latin square design. Only one of the two reciprocal types (B6 females × HR8 males) was carried from G5 to G15 and subsequently utilized in the current study.

Progenitors of wild-derived strains: Details of the origins of wild-derived inbred strains are taken from Beck *et al.* (2000)[41]. Founder mice for the strain Watkins Star Lines A and B (WSA and WSB, respectively) were trapped near the town of Centreville, Maryland by Michael Potter (working at the National Cancer Institute) in 1976. WSA and WSB were selected for dark agouti coat color with white head blaze. In 1986 breeders were sent to Eva M. Eicher at The Jackson Laboratory, where the lines have been maintained since as WSA/EiJ and WSB/EiJ. The LEWES/EiJ strain is descended from wild mice trapped by Potter near Lewes, Delaware in 1981. Breeders were sent to Eicher at the Jackson Laboratory in 1995, where the line has been maintained since. The ZALENDE/EiJ and TIRANO/EiJ inbred strains are descended from mice trapped by Richard D. Sage near the villages of Zalende, Switzerland and Tirano, Italy respectively, in the vicinity of the Poschiavo Valley at the Swiss-Italian border. Mice from Sage's colony were transferred to Potter in 1981. A single breeding pair for each strain was transferred to Eicher at The Jackson Laboratory in 1982. The SPRET/EiJ

inbred strain was derived from wild *Mus spretus* mice trapped near Puerto Real, Cadiz province, Spain by Sage in 1978. The Jackson Laboratory's colony was initiated by Eicher from breeders transferred via Potter in 1983. Frozen tissues from animals in the founder populations were maintained at The Jackson Laboratory by Muriel Davidson until 2014, when they were transferred to the Pardo-Manuel de Villena laboratory at the University of North Carolina at Chapel Hill.

Wild mice: Trapping of wild mice was carried out in concordance with local laws, and either did not require approval or was carried out with the approval of the relevant regulatory bodies (depending on the locality and institution). Specifics of trapping and husbandry are detailed in (JPD, JBS, and FPMV in preparation).

***PCR genotyping at* R2d2**

HR selection lines: To investigate the predicted sweep of the $R2d2^{HC}$ allele in the HR selection lines, we estimated *R2d2* allele frequencies at three generations, one before and two during artificial selection. We genotyped 185 randomly selected individuals from generation -2 and 157 individuals from generation +22 for a marker closely linked to *R2d2*. An additional 80 individuals from generation +61 were genotyped with the MegaMUGA array (see "Microarray genotyping and quality-control" below).

Crude whole-genomic DNA was extracted from mouse tails. The tissues were heated in 100 µl of 25 mM NaOH/0.2 mM EDTA at 95°C for 60 minutes followed by the addition of 100 µl of 40 mM Tris-HCl. The mixture was then centrifuged at 2000 x *g* for 10 minutes and the supernatant used as PCR template.

The *R2d2* element has been mapped to a 900 kb critical region on Chromosome 2: 83,631,096 – 84,541,308 (mm9 build), referred to herein as the "candidate interval"[12]. We designed primers to target a 318 bp region (chr2: 83,673,604 – 83,673,921) within the candidate interval with two distinct haplotypes in linkage with either the $R2d2^{LC}$ allele or the $R2d2^{HC}$ allele. Primers were designed using IDT PrimerQuest (https://www.idtdna.com/Primerquest/Home/Index). Final primer sequences were 5'-CCAGCAGTGATGAGTTGCCATCTTG-3' (forward) and 5'-TGTCACCAAGGTTTTCTTCCAAAGGGAA-3' (reverse).

PCR reactions contained 1 µL dNTPs, 0.3 µL of each primer, 5.3 µL of water, and 0.1 µL of GoTaq polymerase (Promega) in a final volume of 10 µL. Cycling conditions were 95°C, 2-5 min, 35 cycles at 95°, 55° and 72°C for 30 sec each, with a final extension at 72°C, 7 min.

Products were sequenced at the University of North Carolina Genome Analysis Facility on an Applied Biosystems 3730XL Genetic Analyzer. Chromatograms were analyzed with the Sequencher software package (Gene Codes Corporation, Ann Arbor, Michigan, United States).

Assignment to haplotypes was validated by comparing the results to qPCR assays for the single protein-coding gene within *R2d2*, *Cwc22* (see "Copy-number assays" below). For generation +61, haplotypes were assigned based on MegaMUGA genotypes and validated by the normalized per-base read depth from whole-genome sequencing (see below), calculated with samtools mpileup[42]. The concordance between qPCR, read depth, and haplotypes assigned by MegaMUGA or Sanger sequencing is shown in **Extended Data Fig. 7**.

HR8xC57BL/6J advanced intercross line: Tissues were obtained from breeding stock at generations 3, 5, 8, 9, 10, 11, 12, 13, 14 and 15. Crude whole-genomic DNA was extracted by the method described above. We designed primers to amplify a 518 bp region (chr2: 83,724,728 – 83,725,233) within the *R2d2* candidate interval. The amplicon is predicted, based on whole-genome sequencing, to contain a 169 bp deletion in HR8 relative to the C57BL/6J reference genome: 5'-GAGATTTGGATTTGCCATCAA-3' (forward) and 5'-GGTCTACAAGGACTAGAAACAG-3' (reverse). PCR reactions were carried out as described above. Products were visualized and scored on 2% agarose gels.

Whole-genome sequencing of HR selection lines. Ten individuals from generation +61 of each of the eight HR selection lines were subject to whole-genome sequencing. Briefly, high-molecular-weight genomic DNA was extracted using a standard phenol/chloroform procedure. Illumina TruSeq libraries were constructed using 0.5 µg starting material, with fragment sizes between 300 and 500 bp. Each library was sequenced on one lane of an Illumina HiSeq2000 flowcell in a single 2x100bp paired-end run.

Microarray genotyping and quality control. Whole-genomic DNA was isolated from tail, liver, muscle or spleen using Qiagen Gentra Puregene or DNeasy Blood & Tissue kits according to the manufacturer's instructions. All genome-wide genotyping was performed using the Mouse Universal Genotyping Array (MUGA) and its successor, MegaMUGA (GeneSeek, Lincoln, NE)[26,43]. Genotypes were called using Illumina BeadStudio (Illumina Inc., Carlsbad, CA). We excluded all markers and all samples with missingness greater than 10%.

We also computed the sum intensity for each marker: $S_i = X_i + Y_i$, where X_i and Y_i are the normalized hybridization intensities of the two allelic probes. We determined the expected distribution of sum intensity values using a large panel of control samples. We excluded any array for which the set of intensities $I = \{S_1, S_2, \ldots, S_n\}$ was not normally distributed or whose mean was significantly left-shifted with from the reference distribution (one-tailed *t*-test with $p < 0.05$).

Haplotype frequency estimation in the Diversity Outbred. We inferred the haplotypes of DO individuals using probabilistic methods[44,45]. We combined the haplotypes of DO individuals genotyped in this study with the Generation 8 individuals in Didion *et al.* (2015). As an additional QC step, we computed the number of historical recombination breakpoints per individual per generation[20] and removed outliers (more than 1.5 standard deviations from the mean). Next, we excluded related individuals as follows. We used ValBreed[46] to perform a simulation of the DO breeding design for 15 generations to determine the distributions of pairwise haplotype identity between first-degree relatives, second-degree relatives, and unrelated individuals in each generation. We found that all distributions were normal and converged after three generations to mean 0.588 ± 0.045 for first-degree relatives; mean 0.395 ± 0.039 for second-degree relatives; and mean 0.229 ± 0.022 for more distantly related individuals. We then computed the pairwise haplotype identity between all individuals, and identified pairs whose identity had a greater probability of belonging to the first- or second-degree relative distributions than to the unrelated distribution. We iteratively removed the individuals with the greatest number of first- and second-degree

relationships until no related individuals remained. Finally, we computed in each generation the frequency of each founder haplotype at 250 kb intervals surrounding the *R2d2* region (Chromosome 2: 78-86 Mb), and identified the greatest WSB/EiJ haplotype frequency.

Copy-number assays and assignment of* R2d2 *status. Copy-number at *R2d2* was determined by qPCR for *Cwc22*, the single protein-coding gene in the *R2d* repeat unit, as described in detail in Didion *et al.* (2015). Briefly, we used commercially available TaqMan kits (Life Technologies assay numbers Mm00644079_cn and Mm00053048_cn) to measure the copy number of *Cwc22* relative to the reference genes *Tfrc* (cat. no. 4458366, for target Mm00053048_cn) or *Tert* (cat. no. 4458368, for target Mm00644079_cn). Cycle thresholds (C_t) were determined for each target using ABI CopyCaller v2.0 software with default settings, and relative cycle threshold was calculated as

$$\Delta C_t = C_t^{reference} - C_t^{target}$$

We normalized the ΔC_t across batches by fitting a linear mixed model with batch and target-reference pair as random effects.

Estimation of integer diploid copy numbers > ~3 by qPCR is infeasible without many technical and biological replicates, especially in the heterozygous state. We took advantage of *R2d2* diploid copy-number estimates from whole-genome sequencing for the inbred strains C57BL/6J (0), CAST/EiJ (2) and WSB/EiJ (66), and the (WSB/EiJxC57BL/6J)F_1 (33) to establish a threshold for declaring a sample “high-copy.” For each of the two TaqMan target-reference pairs we

calculated the sample mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of the normalized ΔC_t among CAST/EiJ controls and wild *M. m. castaneus* individuals together. We designated as "high-copy" any individual with normalized ΔC_t greater than $\hat{\mu} + 2\hat{\sigma}$ – that is, any individual with approximately > 95% probability of having diploid copy number >2 at *R2d2*. Individuals with high copy number and evidence of local heterozygosity (a heterozygous call at any of the 13 markers in the *R2d2* candidate interval) were declared heterozygous $R2d2^{HC/LC}$, and those with high copy number and no heterozygous calls in the candidate interval were declared homozygous $R2d2^{HC/HC}$.

Exploration of population structure in wild mice. The wild mice used in this study (**Supplementary Table 1**) are a subset of the Wild Mouse Genetic Survey and are characterized in detail elsewhere (JPD, JBS, and FPMV, in preparation). The majority (325 of a total *n* = 500 mice) were trapped at sites across Europe and the Mediterranean basin (**Extended Data Fig. 3A**, upper panel) and in central Maryland and have predominantly *Mus musculus domesticus* ancestry. Additional *M. m. domesticus* populations were sampled from the Farallon Islands near San Francisco, California (20 mice) and Floreana Island in the Galapagos off the coast of Ecuador (15 mice). Of *M. m. domesticus* samples, 245 have the standard mouse karyotype (2*n* = 40) and 226 carry Robertsonian fusion chromosomes (2*n* < 40)[47]. A set of 29 *M. m. castaneus* mice trapped in northern India and Taiwan (**Extended Data Fig. 3A**, lower panel) were included as an outgroup[48].

Scans for signatures of positive selection based on patterns of haplotype-sharing assume that individuals are unrelated. We identified pairs of related individuals using the *IBS2** ratio[49], defined as *HETHET / (HOMHOM + HETHET)*, where *HETHET* and *HOMHOM* are the count of non-missing markers for which both individuals are heterozygous (share two alleles) and homozygous for opposite alleles (share zero alleles), respectively. Pairs with *IBS2** < 0.75 were considered unrelated. Among individuals which were a member of one or more unrelated pairs, we iteratively removed one sample at a time until no related pairs remained, and additionally excluded markers with minor-allele frequency < 0.05 or missingness > 0.10. The resulting dataset contains genotypes for 396 mice at 58,283 markers.

Several of our analyses required that samples be assigned to populations. Because mice in the wild breed in localized demes and disperse only over short distances (on the order of hundreds of meters)[50], it is reasonable to delineate populations on the basis of geography. We assigned samples to populations based on the country in which they were trapped. To confirm that these population labels correspond to natural clusters we performed two exploratory analyses of population structure. First, classical multidimensional scaling (MDS) of autosomal genotypes was performed with PLINK[51] (--mdsplot --autosome). The result is presented in **Extended Data Fig. 3B-C**, in which samples are colored by population. Second, we used TreeMix[52] to generate a population tree allowing for gene flow using the set of unrelated individuals. Autosomal markers were first pruned to reach a set in approximate linkage equilibrium (plink --indep

25 1). TreeMix was run on the resulting set using the *M. m. castaneus* samples as an outgroup and allowing up to 10 gene-flow edges (treemix -root "cas" -k 10). The result is presented in **Extended Data Fig. 3D**. The clustering of samples by population evident by MDS and the absence of long-branch attraction in the population tree together indicate that our choices of population labels are biologically reasonable.

Scans for selection in wild mice. Two complementary statistics, hapFLK[28] and standardized iHS score[27], were used to examine wild-mouse genotypes for signatures of selection surrounding *R2d2*. The hapFLK statistic is a test of differentiation of local haplotype frequencies between hierarchically-structured populations. It can be interpreted as a generalization of Wright's F_{ST} which exploits local LD. Its model for haplotypes is that of fastPHASE[53] and requires a user-specified value for the parameter *K*, the number of local haplotype clusters. We computed hapFLK in the set of unrelated individuals using *M. m. castaneus* samples as an outgroup for *K* = {4, 8, 12, 16, 20, 24, 28, 32} (hapflk --outgroup "cas" -k {K}) and default settings otherwise.

The iHS score (and its allele-frequency-standardized form |iHS|) is a measure of extended haplotype homozygosis on a derived haplotype relative to an ancestral one. For consistency with the hapFLK analysis, we used fastPHASE on the same genotypes over the same range of *K* with 10 random starts and 25 iterations of expectation-maximization (fastphase –K{K} -T10 -C25) to generate phased haplotypes. We then used selscan[54] to compute iHS scores (selscan --ihs) and standardized the scores in 25 equally-sized bins (selscan-norm --bins 25).

Values in the upper tail of the genome-wide distribution of hapFLK or |iHS| represent candidates for regions under selection. We used percentile ranks directly and did not attempt to calculate approximate or empirical *p*-values.

Detection of identity-by-descent (IBD) in wild mice. As an alternative test for selection we computed density of IBD-sharing using the RefinedIBD algorithm of BEAGLE v4.0 (r1399)[55], applying it to the full set of 500 individuals. The haplotype model implemented in BEAGLE uses a tuning parameter (the "scale" parameter) to control model complexity: larger values enforce a more parsimonious model, increasing sensitivity and decreasing computational cost at the expense of accuracy. The authors recommend a value of 2.0 for ~1M SNP arrays in humans. We increased the scale parameter to 5.0 to increase detection power given (a) our much sparser marker set (77,808 SNPs), and (b) the relatively weaker local LD in mouse versus human populations[56]. We trimmed one marker from the ends of candidate IBD segments to reduce edge effects (java -jar beagle.jar ibd=true ibdscale=5 ibdtrim=1). We retained those IBD segments shared between individuals in the set of 396 unrelated mice. In order to limit noise from false-positive IBD segments, we further removed segments with LOD score < 5.0 or width < 0.5 cM.

An empirical IBD-sharing score was computed in 500 kb bins with 250 kb overlap as:

$$f_n = \frac{\sum_n s_{ij} p_{ij}}{w_{ij}}$$

where the sum in the numerator is taken over all IBD segments overlapping bin n and s_{ij} is an indicator variable which takes the value 1 if individuals i,j share a haplotype IBD in bin n and 0 otherwise. The weighting factor w_{ij} is defined as

$$w_{ij} = 0.001 \times \left(\frac{n_a n_b}{W}\right)^{1/2}$$

with

$$W = \max(n_a n_b)$$

where n_a and n_b are the number of unrelated individuals in the population to which individuals i and j belong, respectively. This weighting scheme accounts for the fact that we oversample some geographic regions (for instance, Portugal and Maryland) relative to others. To explore differences in haplotype-sharing within versus between populations we introduce an additional indicator p_{ij}. Within-population sharing is computed by setting p_{ij} = 1 if individuals i,j are drawn from the same population and p_{ij} = 0 otherwise. Between-population sharing is computed by reversing the values of p_{ij}. The result is displayed in **Figure 2**.

Analysis of LD decay in whole-genome sequence from wild mice. We obtained raw sequence reads for 26 unrelated wild mice from[24] (European Nucleotide Archive project accession PRJEB9450; samples listed in **Supplementary Table 4**) and aligned it to the mouse reference genome (GRCm38/mm10 build) using bwa mem with default parameters. SNPs relative to the reference sequence of Chromosome 2 were called using samtools mpileup v0.1.19-44428cd with maximum per-sample depth of 200. Genotype calls with root-mean-square mapping quality < 30 or genotype quality >20 were treated as

missing. Sites were used for phasing if they had a minor-allele count ≥ 2 and at most 2 missing calls. BEAGLE v4.0 (r1399) was used to phase the samples conditional on each other, using 20 iterations for phasing and default settings otherwise (java -jar beagle.jar phasing-its=20). Sites were assigned a genetic position by linear interpolation on the most recent genetic map for the mouse[44,45].

The *R2d2* candidate interval spans positions 83,790,939 – 84,701,151 in the mm10 reference sequence. As the index SNP for *R2d2*^{HC} we chose the SNP with strongest nominal association with *R2d2* copy number (as estimated by Pezer *et al.* (2015)) within 1 kb of the proximal boundary of the candidate interval. That SNP is chr2:83,790,275T>C. The C allele is associated with high copy number and is therefore presumed to be the derived allele. We computed the extended haplotype homozygosity (EHH) statistic[57] in the phased dataset over a 1 Mb window on each side of the index SNP using selscan (selscan --ehh --ehh-win 1000000). The result is presented in **Figure 2B**. Decay of haplotypes away from the index SNP was visualized as a bifurcation diagram (**Figure 2C**) using code adapted from the R package rehh (https://cran.r-project.org/package=rehh).

Estimation of age of* R2d2^{HC} *alleles in wild mice. To obtain a lower bound for the age of *R2d2*^{HC} and its associated haplotype, we used the method of Stephens *et al.* (1998)[58]. Briefly, this method approximates the probability *P* that a haplotype is not broken by recombination during the *G* generations since its origin as

$$P = e^{-G(-\mu+r)}$$

where μ and r are the per-generation rates of mutation and recombination, respectively. Taking P', the observed number of ancestral (non-recombined) haplotypes in a sample, as an estimator of P, obtain the following expression for G:

$$G = -(\log P')/r$$

We enumerated haplotypes in our sample of 52 chromosomes at 3 SNPs spanning the *R2d2* candidate interval. The most proximal SNP is the index SNP for the EHH analyses (chr2:83,790,275T>C); the most distal SNP is the SNP most associated with copy number within 1 kbp of the boundary of the candidate interval (chr2:84,668,280T>C); and the middle SNP was randomly-chosen to fall approximately halfway between (chr2:84,079,970C>T). The three SNPs span genetic distance 0.154 cM (corresponding to r = 0.00154). The most common haplotype among samples with high copy number according to Pezer et al. was assumed to be ancestral. Among 52 chromosomes, 22 carried at least part of the $R2d2^{HC}$-associated haplotype; of those, 11 were ancestral and 11 recombinant (**Supplementary Table 4**). This gives an estimated age of 450 generations for $R2d2^{HC}$.

It should be noted that the approximations underlying this model assume constant population size and neutrality. To the extent that LD decays more slowly on a positively- (or selfishly-) selected haplotype, we will underestimate the true age of $R2d2^{HC}$.

Null simulations of closed breeding populations. Widespread fixation of alleles due to drift is expected in small, closed populations such as the HR lines

or the HR8xC57BL/6J advanced intercross line. But even in these scenarios, an allele under positive selection is expected to fix 1) more often than expected by drift alone in repeated breeding experiments using the same genetic backgrounds, and 2) more rapidly than expected by drift alone. We used the R package simcross (https://github.com/kbroman/simcross) to obtain the null distribution of fixation times and fixation probabilities for an HR line under Mendelian transmission.

We assume that the artificial selection applied for voluntary exercise in the HR lines (described in Swallow *et al.* (1998)) was independent of *R2d2* genotype. This assumption is justified for two reasons. First, 3 of 4 selection lines and 2 of 4 control (unselected) lines fixed $R2d2^{HC}$. Second, at the fourth and tenth generation of the HR8xC57BL/6J advanced intercross, no quantitative trait loci (QTL) associated with the selection criteria (total distance run on days 5 and 6 of a 6-day trial) were found on Chromosome 2. QTL for peak and average running speed were identified at positions linked to *R2d2*; however, HR8 alleles at those QTL were associated with decreased, not increased, running speed[39,40].

Without artificial selection an HR line reduces to an advanced intercross line maintained by avoidance of sib-mating. We therefore simulated 100 replicates of an advanced intercross with 10 breeding pairs and initial focal allele frequency 0.75. Trajectories were followed until the focal allele was fixed or lost. As a validation we confirmed that the focal allele was fixed in 754 of 1000 runs, not different from the expected 750 (p = 0.62, binomial test). Simulated trajectories and the distribution of sojourn times are presented in **Extended Data Fig. 1A-B**.

The HR8xC57BL/6J advanced intercross line was simulated as a standard biparental AIL with initial focal allele frequency of 0.5. Again, 1000 replicates of an AIL with 20 breeding pairs were simulated and trajectories were followed until the focal allele was fixed or lost. The result is presented in **Extended Data Fig. 1C-D**.

Simulations of meiotic drive with unlinked modifiers. To explore the population dynamics of a meiotic drive system in which transmission ratio at a *responder* locus is controlled by genotype at unlinked *modifier* loci, we simulated populations of constant size under the Wright-Fisher model[34]. Each run is characterized by the following parameters: initial frequency of the responder allele (d); initial frequencies of two modifier alleles (f_1, f_2); population size (N); genetic architecture for transmission distortion (additive or epistatic); and effect sizes of modifier alleles (β_1, β_2 for additive model or α for epistatic model). The responder locus and both modifier loci are assumed mutually unlinked. Sex ratio is held constant at 0.5.

At each generation, allele counts at the modifier loci and at the responder locus in males and homozygous females are drawn from a binomial distribution conditional on the previous generation assuming Mendelian segregation (*i.e.,* the standard Wright-Fisher model). Alleles at the responder locus in heterozygous females, however, are drawn by binomial sampling with parameter p conditional on individual genotypes at the modifier loci:

$$p = \begin{cases} 0.5 + x_1\beta_1 + x_2\beta_2 \text{ additive model} \\ 0.5 + x_1 x_2 \alpha \text{ epistatic model} \end{cases}$$

where the x_i are minor-allele counts at the modifier loci.

A run stops when either (a) the responder allele is fixed or lost; or (b) more than 3τ generations have elapsed, where τ is the diffusional approximation to the expected time to fixation or loss of a neutral allele[59]:

$$\tau = -2N\{d \log d \ + \ (1-d)\log(1-d)\}$$

We simulated 100 runs for each possible parameter combination across the following ranges:

$$N = \{10^2, 10^3, 10^4\}$$

$$d = \{0.01, 0.05, 0.10, 0.25, 0.5\}$$

$$f_1 = \{0.01, 0.05, 0.10, 0.25, 0.5, 0.75\}$$

$$f_2 = \{0.01, 0.05, 0.10, 0.25\}$$

$$\beta_i = \{0.0250, 0.0625, 0.1000\}$$

$$\alpha = \{0.10, 0.25, 0.40\}$$

Effect sizes for modifier loci were chosen so that the maximum achievable transmission ratio (at the responder locus) in the population would be 0.60, 0.75 or 0.90. Simulations are summarized in **Extended Data Fig. 6**.

Haplotype analysis around R2d2 in laboratory strains. We used the Mouse Phylogeny Viewer (http://msub.csbio.unc.edu/)[48] to investigate the extent of haplotype-sharing around *R2d2* in inbred strains of *M. musculus*. First we identified the largest interval containing *R2d2* within which the classical inbred strains carrying $R2d2^{HC}$ alleles (ALR/LtJ, ALS/LtJ, CHMU/LeJ, NU/J) all have the

same phylogenetic history: this core interval is Chr2: 82,284,942 – 84,870,179. (Note that individual pairs within that set, *eg*. CHMU/LeJ and ALS/LtJ, share over a longer region.) Next we obtained genotypes for the region Chr2: 75 – 90 Mb from the Mouse Diversity Array (http://cgd.jax.org/datasets/diversityarray/CELfiles.shtml) for the other four classical inbred strains plus other inbred strains with $R2d2^{HC}$ alleles: the selection line HR8 and wild-derived strains RBA/DnJ, RBB/DnJ, RBF/DnJ, WSB/EiJ and ZALENDE/EiJ. We treated WSB/EiJ as the template haplotype and recoded genotypes at each of 2,956 markers as 0, 1 or 2 according to the number of alleles shared with WSB/EiJ. Haplotype-sharing among the wild-derived strains was then assessed by manual inspection. Since the classical inbred strains share a single ancestral haplotype in the core region, and that haplotype is identical to WSB/EiJ, it follows that the wild-derived strains identical to WSB/EiJ share the same haplotype. The result is shown in **Extended Data Fig. 5**.

References

1. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* **74,** 1111–1120 (2004).
2. Williamson, S. H. *et al.* Localizing recent adaptive evolution in the human genome. *PLoS Genet* **3,** e90 (2007).
3. Presgraves, D. C., Gérard, P. R., Cherukuri, A. & Lyttle, T. W. Large-scale selective sweep among segregation distorter chromosomes in African populations of Drosophila melanogaster. *PLoS Genet* **5,** e1000463 (2009).
4. Staubach, F. *et al.* Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). *PLoS Genet* **8,** e1002891 (2012).
5. Grossman, S. R. *et al.* Identifying recent adaptations in large-scale genomic data. *Cell* **152,** 703–713 (2013).
6. Colonna, V. *et al.* Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biol* **15,** R88 (2014).
7. Sandler, L. & Novitski, E. Meiotic drive as an evolutionary force. *American Naturalist* 105–110 (1957).
8. White, M. J. D. *Modes of Speciation*. (W.H.Freeman & Co Ltd, 1978).
9. Henikoff, S. & Malik, H. S. Centromeres: selfish drivers. *Nature* **417,** 227–227 (2002).
10. Pardo-Manuel de Villena, F. in *Mammalian Genomics* (eds. Ruvinsky, A. & Graves, J. A. M.) 317–348 (CABI, 2004).
11. Derome, N., Métayer, K., Montchamp-Moreau, C. & Veuille, M. Signature of selective sweep associated with the evolution of sex-ratio drive in *Drosophila simulans*. *Genetics* **166,** 1357–1366 (2004).
12. Didion, J. P. *et al.* A multi-megabase copy number gain causes maternal transmission ratio distortion on mouse chromosome 2. *PLoS Genet* **11,** e1004850 (2015).
13. Garland, T. & Rose, M. R. *Experimental Evolution*. (2009).
14. Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nat Rev Genet* **14,** 827–839 (2013).
15. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet Res* **23,** 23–35 (1974).
16. Kaplan, N. L., Hudson, R. R. & Langley, C. H. The ‘hitchhiking effect’ revisited. *Genetics* **123,** 887–899 (1989).
17. Pelz, H.-J. *et al.* The genetic basis of resistance to anticoagulants in rodents. *Genetics* **170,** 1839–1847 (2005).
18. Brandvain, Y. & Coop, G. Scrambling eggs: meiotic drive and the evolution of female recombination rates. *Genetics* **190,** 709–723 (2011).
19. Axelsson, E. *et al.* Segregation distortion in chicken and the evolutionary consequences of female meiotic drive in birds. *Heredity* **105,** 290–298 (2010).
20. Svenson, K. L. *et al.* High-resolution genetic mapping using the Mouse Diversity outbred population. *Genetics* **190,** 437–447 (2012).

21. Swallow, J. G., Carter, P. A. & Garland, T., Jr. Artificial selection for increased wheel-running behavior in house mice. *Behavior Genetics* **28,** 227–237 (1998).
22. Kelly, S. A. *et al.* Parent-of-origin effects on voluntary exercise levels and body composition in mice. *Physiol. Genomics* **40,** 111–120 (2010).
23. Didion, J. P. & Pardo-Manuel de Villena, F. Deconstructing *Mus gemischus*: advances in understanding ancestry, structure, and variation in the genome of the laboratory mouse. *Mamm Genome* **24,** 1–20 (2013).
24. Pezer, Ž., Harr, B., Teschke, M., Babiker, H. & Tautz, D. Divergence patterns of genic copy number variation in natural populations of the house mouse (*Mus musculus domesticus*) reveal three conserved genes with major population-specific expansions. *Genome Res* (2015). doi:10.1101/gr.187187.114
25. Rogala, A. R. *et al.* The Collaborative Cross as a resource for modeling human disease: CC011/Unc, a new mouse model for spontaneous colitis. *Mamm Genome* **25,** 95–108 (2014).
26. Morgan, A. P. & Welsh, C. E. Informatics resources for the Collaborative Cross and related mouse populations. *Mamm Genome* (2015). doi:10.1007/s00335-015-9581-z
27. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol* **4,** e72 (2006).
28. Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M. & Servin, B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* **193,** 929–941 (2013).
29. Messer, P. W. & Petrov, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol Evol* **28,** 659–669 (2013).
30. Albrechtsen, A., Moltke, I. & Nielsen, R. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* **186,** 295–308 (2010).
31. Song, Y. *et al.* Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Curr Biol* **21,** 1296–1301 (2011).
32. Nachman, M. W., Boyer, S. N., Searle, J. B. & Aquadro, C. F. Mitochondrial DNA variation and the evolution of Robertsonian chromosomal races of house mice, *Mus domesticus*. *Genetics* **136,** 1105–1120 (1994).
33. Hedrick, P. W. The establishment of chromosomal variants. *Evolution* **35,** 322–332 (1981).
34. Gillespie, J. H. *Population genetics: a concise guide*. (JHU Press, 2010).
35. Hernandez, R. D. *et al.* Classic selective sweeps were rare in recent human evolution. *Science* **331,** 920–924 (2011).
36. Pardo-Manuel de Villena, F. & Sapienza, C. Nonrandom segregation during meiosis: the unfairness of females. *Mamm Genome* **12,** 331–339 (2001).
37. Lyon, M. F. The genetic basis of transmission-ratio distortion and male sterility due to the t complex. *The American Naturalist* **137,** 349–358

(1991).
38. Hartl, D. L. Complementation analysis of male fertility among the segregation distorter chromosomes of *Drosophila melanogaster*. *Genetics* **73,** 613–629 (1973).
39. Kelly, S. A. *et al.* Genetic architecture of voluntary exercise in an advanced intercross line of mice. *Physiol. Genomics* **42,** 190–200 (2010).
40. Leamy, L. J., Kelly, S. A., Hua, K. & Pomp, D. Exercise and diet affect quantitative trait loci for body weight and composition traits in an advanced intercross population of mice. *Physiol. Genomics* **44,** 1141–1153 (2012).
41. Beck, J. A. *et al.* Genealogies of mouse inbred strains. *Nat Genet* **24,** 23–25 (2000).
42. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
43. Collaborative Cross Consortium. The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics* **190,** 389–401 (2012).
44. Liu, E. Y., Zhang, Q., McMillan, L., Pardo-Manuel de Villena, F. & Wang, W. Efficient genome ancestry inference in complex pedigrees with inbreeding. *Bioinformatics* **26,** i199–207 (2010).
45. Liu, E. Y. *et al.* High-resolution sex-specific linkage maps of the mouse reveal polarized distribution of crossovers in male germline. *Genetics* **197,** 91–106 (2014).
46. Kover, P. X. *et al.* A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genet* **5,** e1000551 (2009).
47. Piálek, J., Hauffe, H. C. & Searle, J. B. Chromosomal variation in the house mouse. *Biological Journal of the Linnean Society* **84,** 535–563 (2005).
48. Yang, H. *et al.* Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet* **43,** 648–655 (2011).
49. Stevens, E. L. *et al.* Inference of relationships in population data using identity-by-descent and identity-by-state. *PLoS Genet* **7,** e1002287 (2011).
50. Pocock, M., Hauffe, H. C. & Searle, J. B. Dispersal in house mice. *Biological Journal of the Linnean Society* **84,** 565–583 (2005).
51. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81,** 559–575 (2007).
52. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8,** e1002967 (2012).
53. Scheet, P. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78,** 629–644 (2006).
54. Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol* **31,** 2824–2827 (2014).
55. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194,** 459–471

(2013).
56. Laurie, C. C. *et al.* Linkage disequilibrium in wild mice. *PLoS Genet* **3,** e144 (2007).
57. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).
58. Stephens, J. C. *et al.* Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *Am J Hum Genet* **62,** 1507–1515 (1998).
59. Kimura, M. & Ohta, T. The average number of generations until fixation of a mutant gene in a finite population. *Genetics* **61,** 763–771 (1969).

Acknowledgements

We wish to thank all the scientists and research personnel who collected and processed the samples used in this study. In particular we acknowledge Luanne Peters and Alex Hong-Tsen Yu for providing critical samples; Ryan Buus and T. Justin Gooch for isolating DNA for high-density genotyping of wild-caught mice; and Vicki Cappa, A.Cerveira, Daniel Förster, G. Ganem, Ron and Annabelle Lesher, K. Saïd, Toni Schelts, Dan Small, and J. Tapisso for aiding in mouse trapping. We thank Muriel Davisson at the Jackson Laboratory for maintaining, for several decades, tissue samples from breeding colonies used to derive wild-derived inbred strains. This work was funded in part by T32GM067553 (JPD, APM), F30MH103925 (APM); Vaadia-BARD Postdoctoral Fellowship Award FI-478-13 (LY); W81XWH-11-1-0762 (CJB); University of Rome "La Sapienza" (RC, ES); The Jackson Laboratory new investigator funds (EJC); P50GM076468 (EJC, GAC and FPMV); K01MH094406 (JJC); CNRS (JBD); NSF IOS-1121273 (TG); Claraz-Stiftung (AL); PRDC/BIA-EVF/116884/2010 and UID/AMB/50017/2013 (MM, JBS); the intramural research program of NIDDK, NIH (BR and SPR); DK-076050, DK-056350 (DP); and a grant to The Jackson Laboratory's Nathan Shock Center, AG038070 (GAC), and the Oliver Smithies Investigator funds provided by the School of Medicine at University of North Carolina (FPMV).

Author Contributions

JPD, GAC and FPMV conceived the study. JBD, CJB, KJC, RC, Y-HC, AJC, JJC, EJC, JEF, SIG, DMG, TG, EBG-A, MDG, SAG, IG, AH, HCH, JSH, JMH, KH, WJJ, AKL, MJL-F, GM, MM, LM, MGR, BR, SPR, JBS, MSS, ES, KLS, PT-L, DWT, JVQ, GMW, DP, GAC, and FPMV provided biological samples and/or unpublished data sets; APM, LY, TAB, RCM, and LOdS conducted experiments. JPD, APM, and LY analyzed the data. JPD, APM, and FPMV wrote the paper.

Author Information

All data is made available at http://csbio.unc.edu/r2d2/. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to FPMV (fernando@med.unc.edu).

Figure Legends

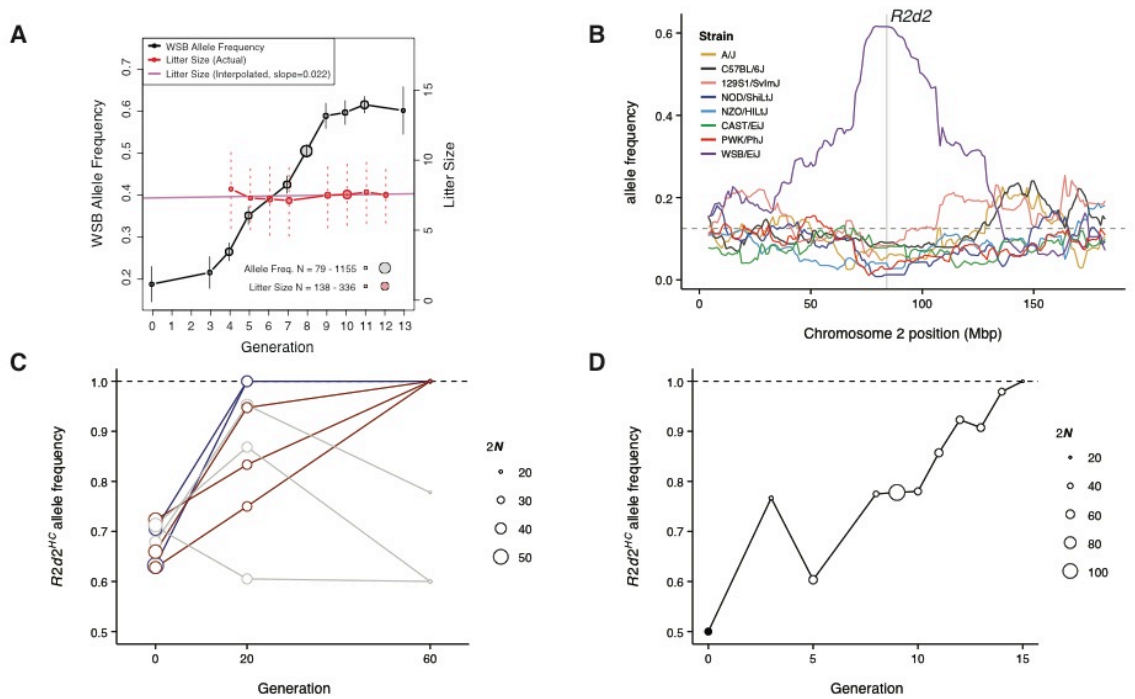


Figure 1. *R2d2*^{HC} alleles rapidly increase in frequency in three laboratory populations. (A) *R2d2* drives three-fold increase in WSB/EiJ allele frequency in 13 generations in the DO population. WSB/EiJ allele frequency (black circles, left y-axis) and mean first litter size (red circles, right y-axis) measured in cohorts ("generations") with available data. Circle sizes reflect numbers of individuals (black) and litters (red); vertical lines: standard error; pink line: linear interpolation of litter size. **(B)** Allele frequencies across Chromosome 2 (averaged in 1 Mb bins) at generation 13 of the DO, classified by founder strain. Grey shaded region is the candidate interval for *R2d2*. **(C)** *R2d2*^{HC} allele frequency during breeding of 4 HR selection lines and 4 control lines. Trajectories are colored by their fate: blue, *R2d2*^{HC} fixed by generation 20; red. *R2d2*^{HC} fixed by generation 60; grey, *R2d2*^{HC} not fixed. Circle sizes reflect number of chromosomes (*2N*) genotyped.

(D) *R2d2*^{HC} allele frequency during breeding of an (HR8xC57BL/6J) advanced intercross line. Circle sizes reflect number of chromosomes (*2N*) genotyped.

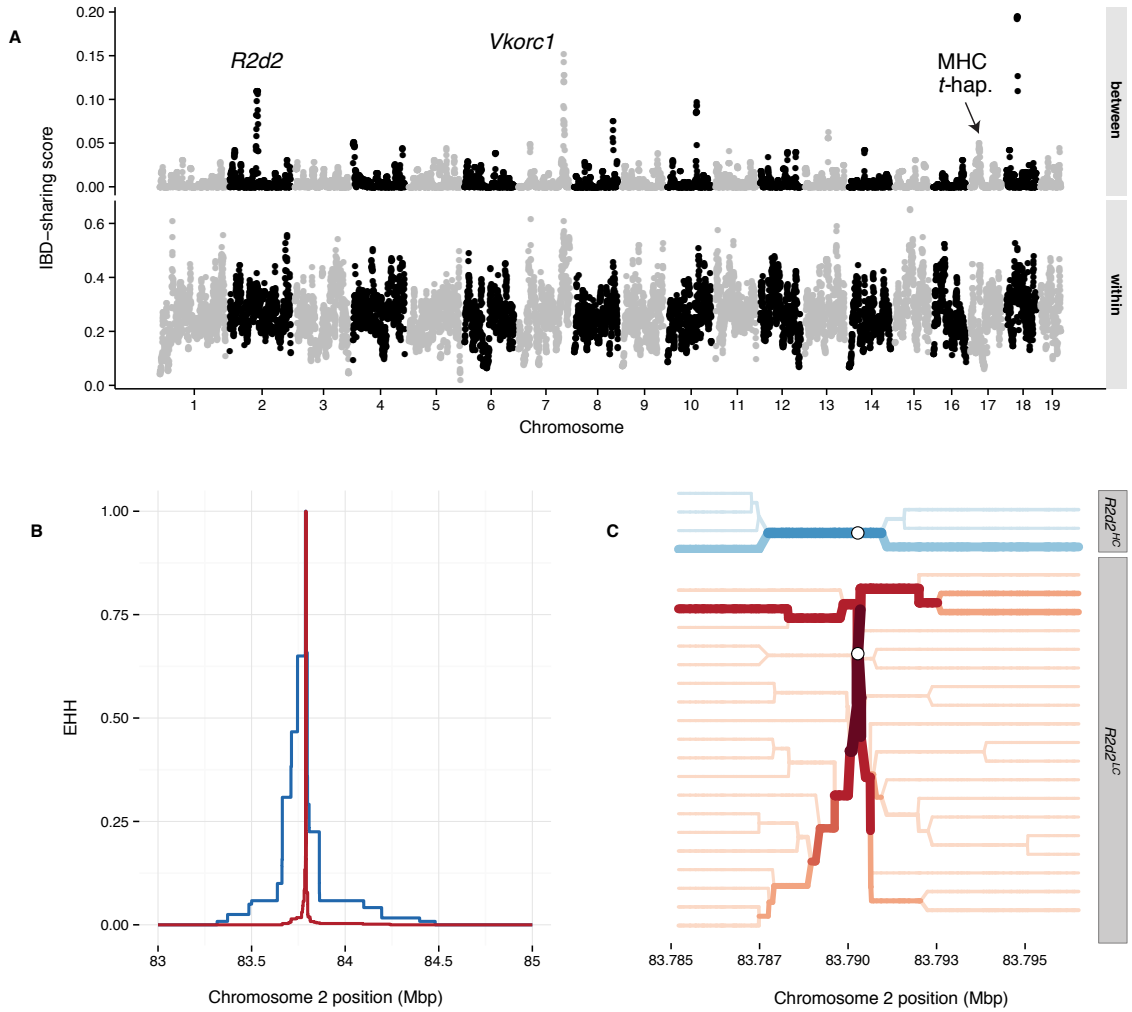
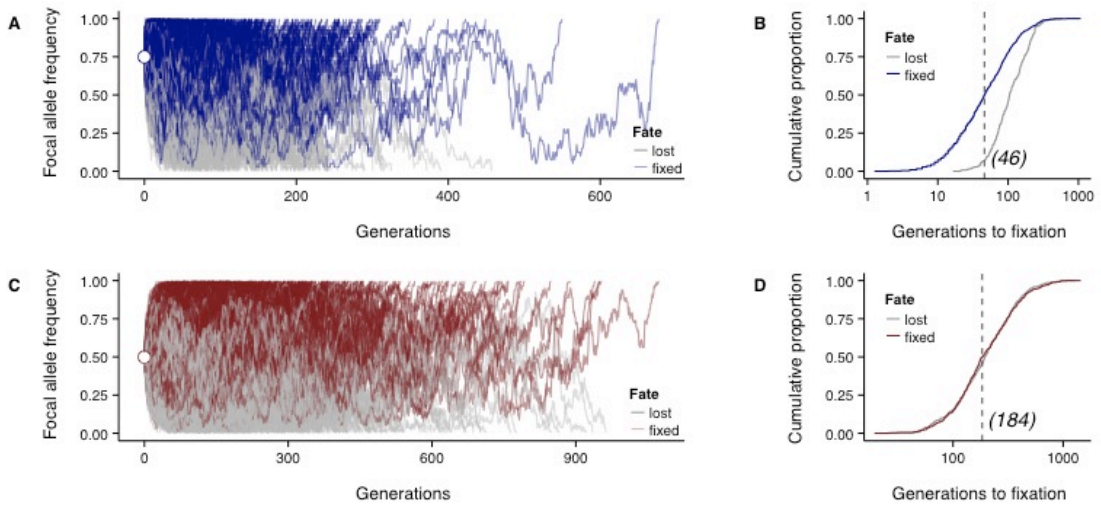


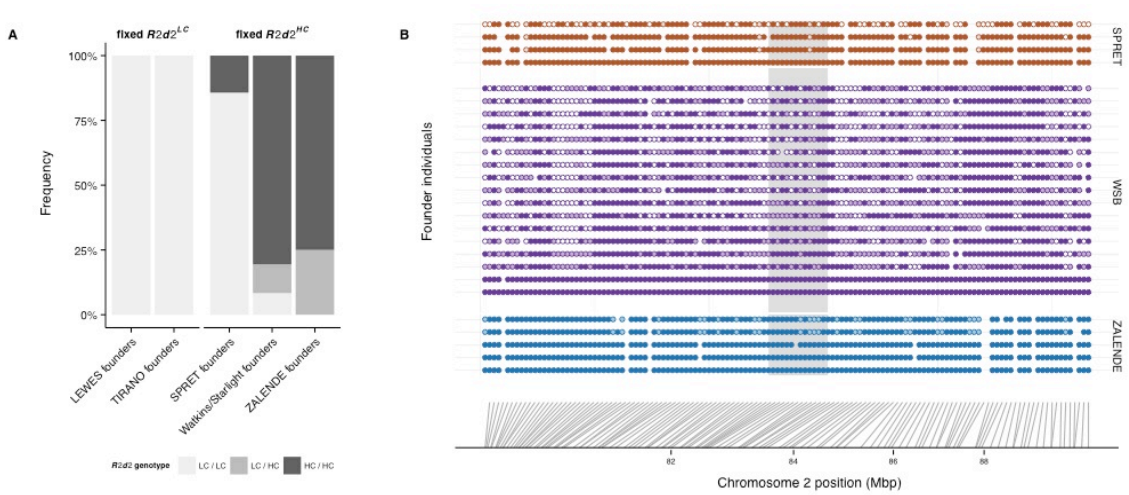
Figure 2. Haplotype-sharing at *R2d2* provides evidence of a selective sweep in wild mice of European origin. (A) Weighted haplotype-sharing score (see **Methods**), computed in 500 kb bins across autosomes, when those individuals are drawn from the same population (lower panel) or different populations (upper panel). Peaks of interest overly *R2d2* (Chromosome 2), *Vkorc1* (distal Chromosome 7). The position of the closely-linked *t*-haplotype and MHC loci is also marked. **(B)** Decay of extended haplotype homozygosity

(EHH)[57] on the *R2d2*^{HC}-associated (blue) versus the *R2d2*^{LC}-associated (red) haplotype. EHH is measured outward from the index SNP at chr2:83,790,275 and is bounded between 0 and 1. **(C)** Haplotype bifurcation diagrams for the *R2d2*^{HC} (top panel, red) and *R2d2*^{LC} (bottom panel, blue) haplotypes at the index SNP (open circle). Darker colors and thicker lines indicate higher haplotype frequencies. Haplotypes extend 100 sites in each direction from the index SNP.

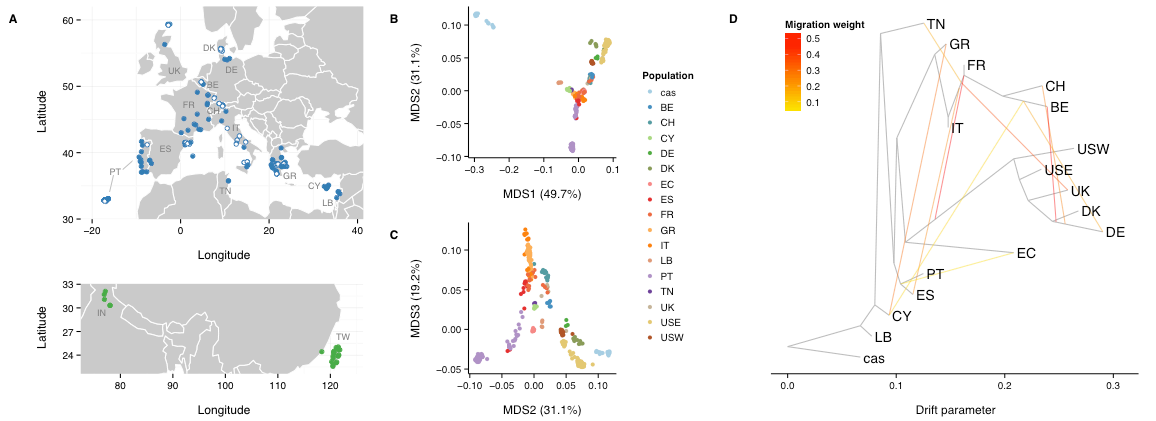
Extended Data Figure Legends



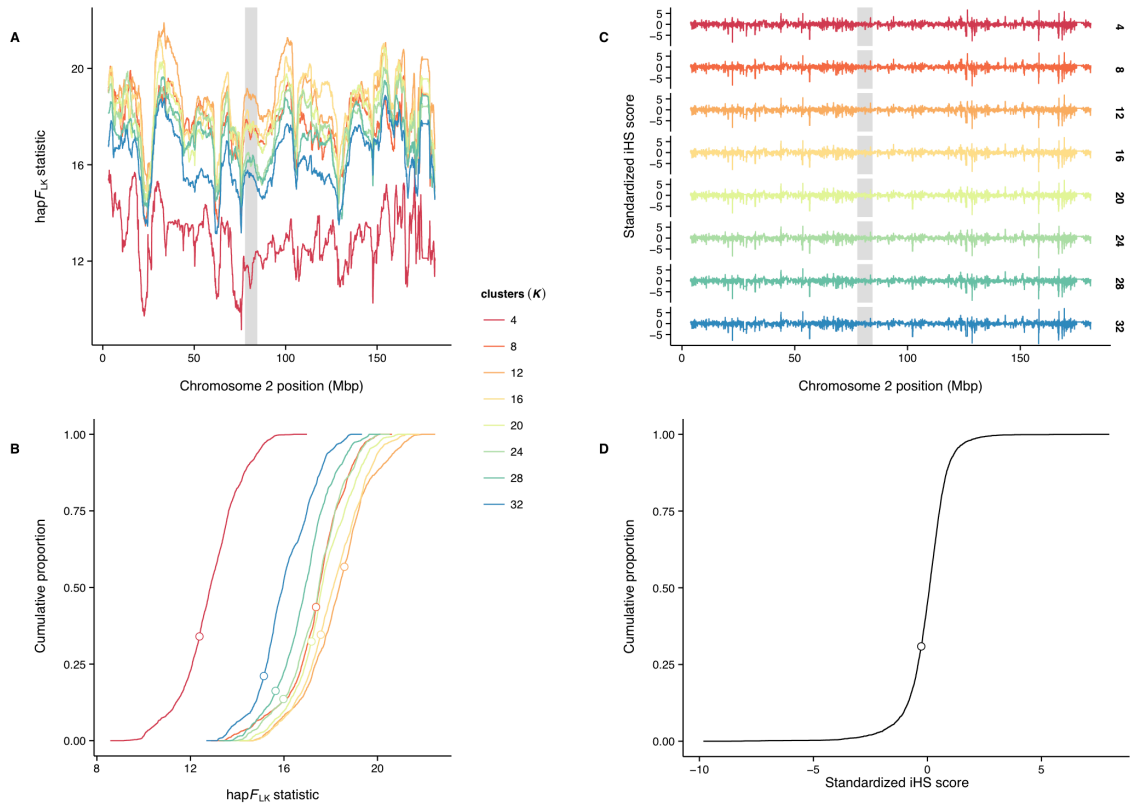
Extended Data Fig. 1. Simulations of closed breeding populations under Mendelian segregation. (A) Frequency trajectories of focal allele in 1,000 simulations of an intercross line mimicking the HR breeding scheme, colored according to fate (blue if focal allele fixed; grey if lost). Open circle indicates initial frequency of the focal allele. **(B)** Cumulative distribution of time to fixation (blue) or loss (grey) of the focal allele. Dotted line indicates median fixation time. **(C)** Frequency trajectories of focal allele in 1,000 simulations of an advanced intercross line mimicking the HR8xC57BL/6J AIL, colored according to fate (red if focal allele fixed; grey if lost). Open circle indicates initial frequency of the focal allele. **(D)** Cumulative distribution of time to fixation (blue) or loss (grey) of the focal allele. Dotted line indicates median fixation time.



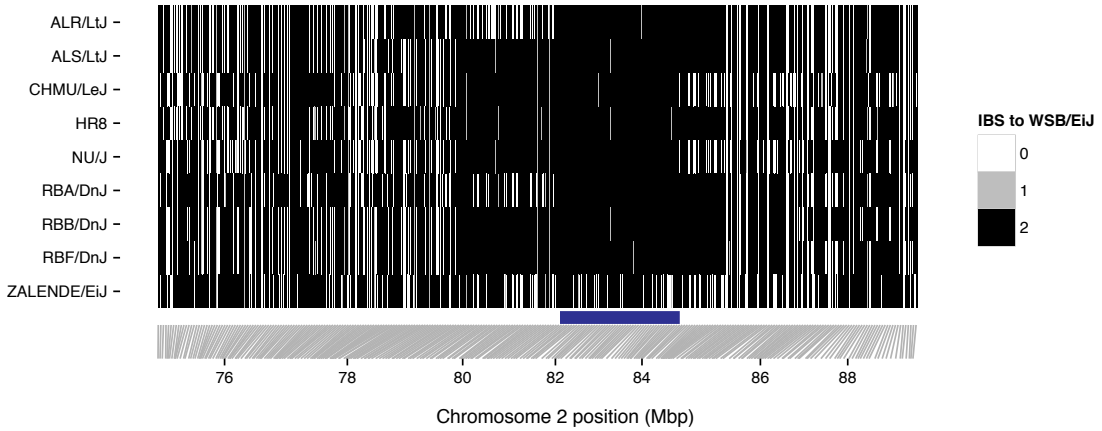
Extended Data Fig. 2. Multiple wild-derived inbred lines have fixed *R2d2^{HC}* alleles that were segregating in founder populations. (A) *R2d2* genotype frequencies in available ancestors of wild-derived inbred lines, determined by qPCR (see **Methods** and **Extended Data Fig. 7**). **(B)** Genotypes at markers on the MegaMUGA array (see **Methods**) in the region Chromosome 2: 80 Mb – 90 Mb for founder individuals of the SPRE/EiJ (brown) or ZALENDE/EiJ (blue) inbred lines. For WSB/EiJ (purple), genotypes are from present-day wild individuals from the township of Centreville, Maryland. Genotypes are coded by identity-by-state (IBS) to the respective inbred line: dark circles, homozygous for allele fixed in inbred line; light circles, heterozygous; open circles, homozygous for alternative allele. *R2d2* candidate region is indicated by grey shaded region. This panel demonstrates that the founders of each line were most likely segregating for *R2d2*.



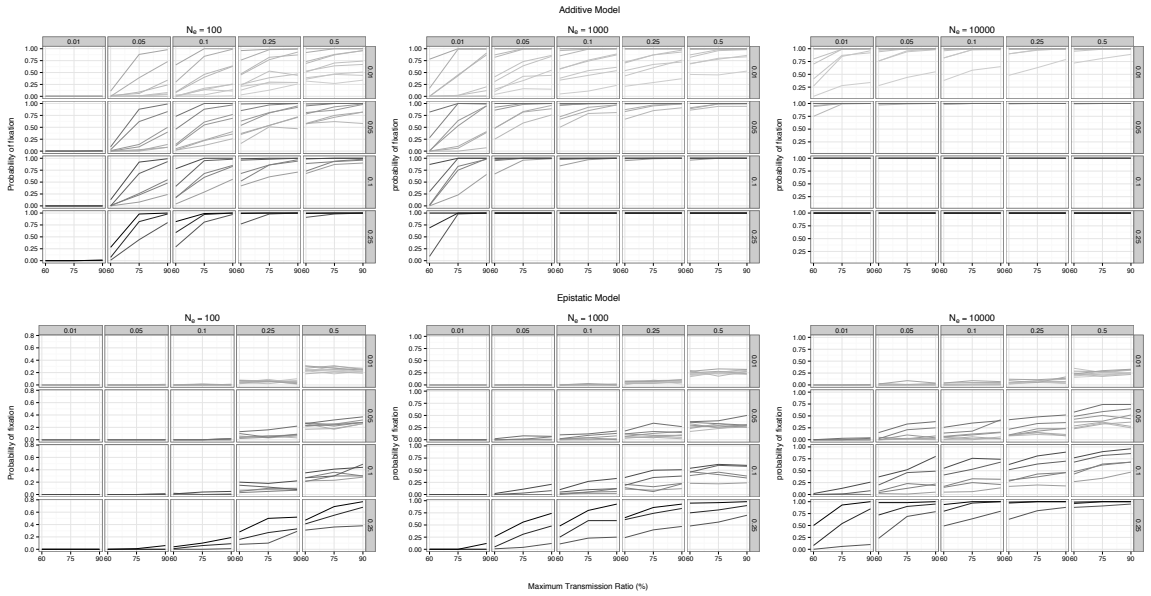
Extended Data Fig. 3. Wild mouse populations used in this study. (A) Geographic distribution of samples used in this study. Samples are colored by taxonomic origin: blue for *M. m. domesticus*, green for *M. m. castaneus*. Those with standard karyotype (2*n* = 40) are indicated by closed circles; samples with Robertsonian fusion karyotypes (2*n* < 40) are indicated by open circles. Populations from Floreana Island (Galapagos Islands, Ecuador; "EC"), Farallon Island (off the coast of San Francisco, California, United States; "USW"), and Maryland, United States ("USE") are not shown. **(B,C)** Multidimensional scaling (MDS) (*k* = 3 dimensions) reveals population stratification consistent with geography. *M. m. domesticus* populations are labeled by country of origin. Outgroup samples of *M. m. castaneus* origin are combined into a single cluster ("cas"). **(D)** Population graph estimated from autosomal allele frequencies by TreeMix. Black edges indicate ancestry, while red edges indicate gene flow by migration or admixture. Topography of the population graph is consistent with MDS result and with the geographic origins of the samples.



Extended Data Fig. 4. Tests for selection based on population differentiation and haplotype length do not detect sweeps at *R2d2*. (A) Plot of hapFLK statistic along Chromosome 2, for a range of values of the model parameter *K* (number of local haplotype clusters). **(B)** Cumulative distribution of hapFLK across autosomes, for a range of values of *K*. Value of the statistic at *R2d2* is indicated by open circle. **(C)** Plot of standardized iHS score along Chromosome 2 after phasing with fastPHASE, for a range of values of *K*. **(D)** Cumulative distribution of standardized iHS scores across autosomes after fastPHASE with *K* = 12. Value of the statistic at *R2d2* is indicated by open circle.

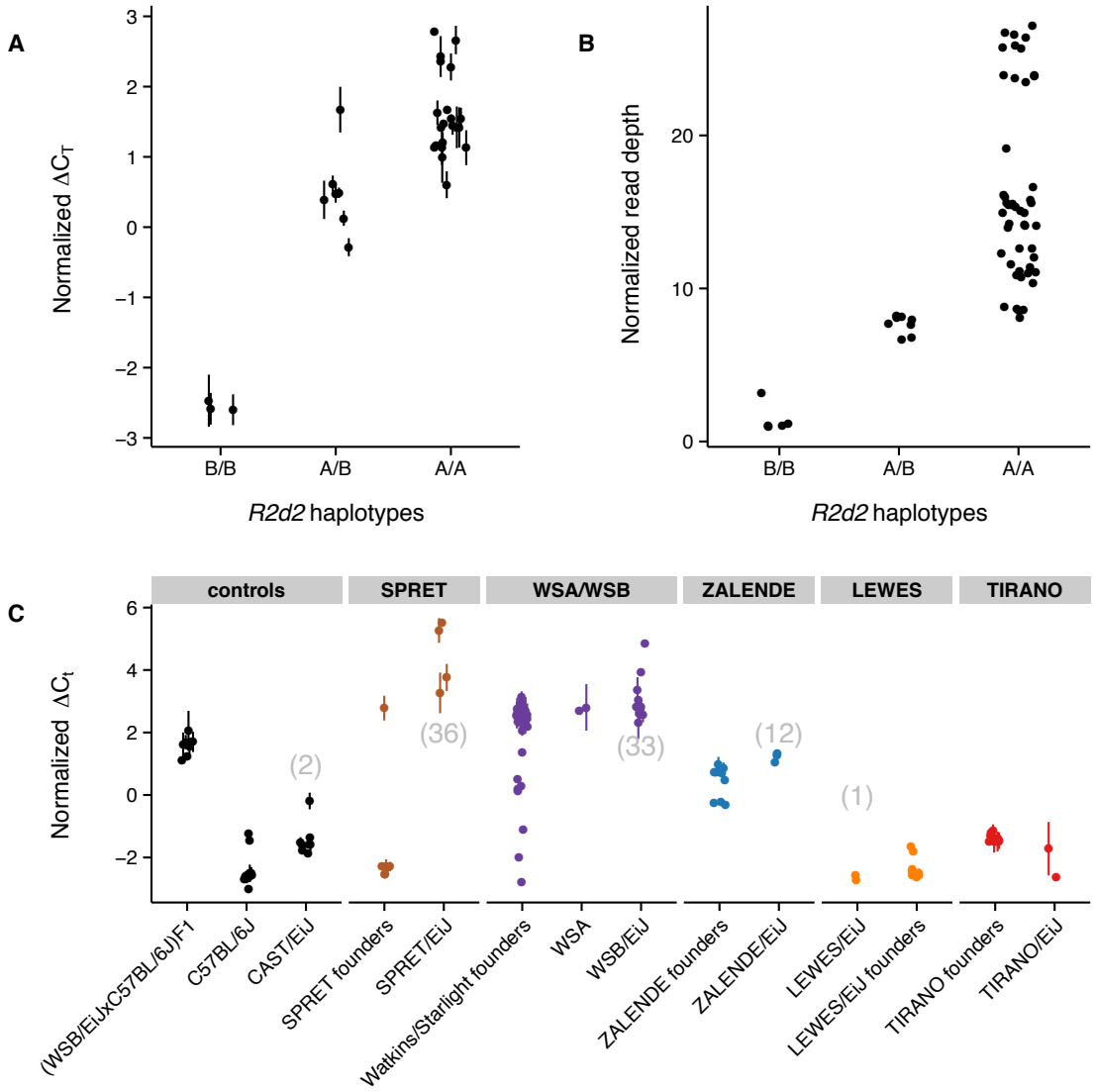


Extended Data Fig. 5. Most inbred strains carrying *R2d2*^{HC} share a haplotype over 2 – 5 Mb. Genotypes from the Mouse Diversity Array at markers in the region Chr 2: 75 – 90 Mb, coded by identity-by-state (IBS) to WSB/EiJ: black, homozygous for WSB/EiJ allele; grey, heterozygous; white, homozygous for alternative allele. All inbred strains with *R2d2*^{HC} alleles except ZALENDE/EiJ share a core 2.2 Mb haplotype (blue bar) with WSB/EiJ.



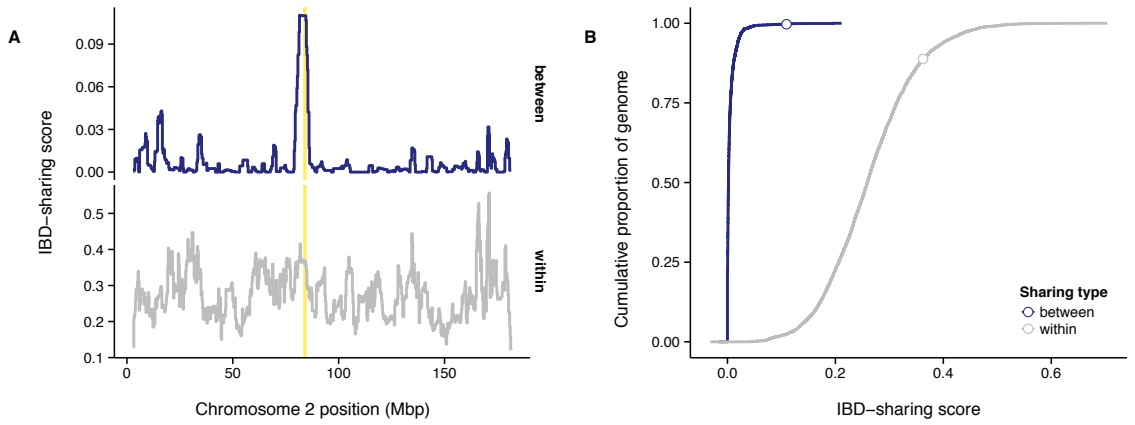
Extended Data Fig. 6. Forward-in-time simulations of a meiotic drive system with two unlinked modifier alleles. Y-axis gives proportion of 100 simulation runs that result in fixation of responder allele within 3τ generations,

where τ is the expected time to fixation or loss of a neutral allele (see Methods). X-axis gives effect size of modifier loci in terms of the maximum achievable transmission ratio (where normal Mendelian ratio is 50%). Sub-panels are defined by initial frequency of the responder allele (across columns) and initial frequency of the rarer of the two modifier alleles (across rows). Lines are colored by joint allele frequency at modifier loci, with darker shades indicating higher frequency. Two genetic architectures were simulated for meiotic drive: additive effects of genotype at modifier loci in the top row, with population sizes (A) 100, (B) 1000 and (C) 10000; or epistatic effects in the bottom row, with population sizes (D) 100, (E) 1000, and (F) 10000.



Extended Data Fig. 7. Characterization of *Cwc22* qPCR assays. (A) Concordance between local haplotype and qPCR in HR lines. Normalized ΔC_t from qPCR assay against *Cwc22* versus local haplotype at Chromosome 2: 83 Mb (A = $R2d2^{LC}$, B = $R2d2^{HC}$) in HR generation +61 individuals. Error bars represent mean +/- 1 SD over technical replicates, when present. **(B)** Normalized read depth at *R2d2* in whole-genome sequencing versus local haplotype. (**C**) *R2d2* copy number of wild-derived inbred mouse lines and available ancestors, estimated by qPCR. Samples listed as "control" are included as internal

calibration points. For inbred strains that have been sequenced (CAST/EiJ, SPRET/EiJ, WSB/EiJ, ZALENDE/EiJ, LEWES/EiJ) copy numbers estimated from depth of coverage are indicated in parentheses.



Extended Data Fig. 8. Haplotype-sharing on Chromosome 2 among wild mice of European origin. (A) Weighted haplotype-sharing score (see **Methods**), computed in 500 kb bins across Chromosome 2, when those individuals are drawn from the same population (grey line, lower panel) or different populations (blue line, upper panel). Candidate interval for *R2d2* is indicated by yellow shaded region. This panel is a magnified view of **Figure 2A**. **(B)** Cumulative distribution of IBD-sharing probability across all autosomes either within (grey line) or between (blue line) populations. Open circles indicate value at *R2d2*.

Supplementary Tables

Supplementary Table 1. Wild mice used in this study (*n* = 500). Column legend is as follows. **Taxon**: "Cas" (*M. m. castaneus*), "Dom" (*M. m. domesticus*). **Countries** are denoted using ISO 2166 standard 2-letter codes. (**Chromosomal) Races** follow the nomenclature of[47]. **TaqMan mean, SD**: mean and standard deviation of normalized ΔC_t from qPCR assay(s) performed on this sample. **TaqMan target**: 1 (assay Mm00053048_cn) or 2 (assay Mm00644079_cn). ***R2d2* zygosity**: "het", sample is heterozygous at one or more markers in the *R2d2* candidate interval; "hom", sample is homozygous at all markers in the *R2d2* candidate interval. ***R2d2* genotype**: coded as number of chromosomes with an $R2d2^{HC}$ allele in this sample (0, 1 or 2). **Unrelated**: TRUE if this sample is a member of the subset of 396 unrelated mice. NA: data not available.

Supplementary Table 2. $R2d2^{HC}$ allele frequencies in wild *M. m. domesticus* populations.

Supplementary Table 3. Diversity outbred mice used to determine *R2d2* allele frequencies.

Supplementary Table 4. Wild *M. m. domesticus* samples from Pezer *et al.* (2015) (*n* = 26). Column key is as follows. **Name**: sample name in this study. **Old name:** sample name used in Pezer et al. (2015). Locality: approximate trapping locations within indicated countries. ***Cwc22* copy number**: Estimated diploid copy number of *Cwc22*, rounded to nearest integer, as reported in Supplementary Table 4 of Pezer et al. (2015). ***R2d2* copy number**: copy-number classification at *R2d2*, using *Cwc22* as proxy; "high" when >2, "low" otherwise.

1 ***R2d2* hap 1**: inferred first haplotype at 3 SNPs across *R2d2* candidate interval;

2 alleles are coded as 1 = *R2d2*^{LC}-associated, 2 = *R2d2*^{HC}-associated. ***R2d2* hap**

3 **2**: inferred second haplotype across *R2d2*.