High-Resolution Maps of Mouse Reference Populations

Petr Simecek[1,2], Jiri Forejt[2], Robert W. Williams[3], Toshihiko Shiroishi[4], Toyoyuki Takada[4], Lu Lu[3], Thomas E. Johnson[5], Beth Bennett[5], Christian F. Deschepper[6], Marie-Pier Scott-Boyer[6], Gary Churchill[1,8], Fernando Pardo-Manuel de Villena[7,8]

1) The Jackson Laboratory, Bar Harbor, ME, US

2) Institute of Molecular Genetics of the ASCR, Division BIOCEV, Vestec, Czech Rep.

3) University of Tennessee Health Science Center, Memphis, TN, US

4) National Institute of Genetics, Japan

5) University of Colorado at Boulder, CO, US

6) Institut de Recherches Cliniques, Montreal, QC, Canada

7) Department of Genetics, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, US

8) Corresponding authors: Gary Churchill (Gary.Churchill@jax.org) and Fernando Pardo-Manuel de Villena (Fernando@med.unc.edu)

Running title: High-Resolution Maps of Reference Populations

Keywords: chromosome substitution strains, recombinant inbred strains, mouse diversity genotyping array, gene conversions

Correspondence to:

Gary Churchill
The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609 USA
Tel: +1 207-288-6189
Email: gary.churchill@jax.org

Fernando Pardo-Manuel de Villena
University of North Carolina at Chapel Hill, 5046 Genetic Medicine Bldg, CB#7264,
Chapel Hill, NC 27599-7264 USA
Tel: +1 919-843-5403
Email: fernando@med.unc.edu

Abstract

Genetic reference panels are widely used to map complex, quantitative traits in model organisms. We have generated new high-resolution genetic maps of 259 mouse inbred strains from recombinant inbred strain panels (C57BL/6J x DBA/2J, ILS/IbgTejJ x ISS/IbgTejJ, C57BL/6J x A/J) and chromosome substitution strain panels (C57BL/6J-Chr#<A/J>, C57BL/6J-Chr#<PWD/Ph>, C57BL/6J-Chr#<MSM/Ms>). We genotyped all samples using the Affymetrix Mouse Diversity Array with an average inter-marker spacing of 4.3kb. The new genetic maps provide increased precision in the localization of recombination breakpoints compared to the previous maps. Although the strains were presumed to be fully inbred, we found residual heterozygosity in 40% of individual mice from five of the six panels. We also identified *de novo* deletions and duplications, in homozygous or heterozygous state, ranging in size from 21kb to 8.4Mb. Almost two–thirds (46 out of 76) of these deletions overlap exons of protein coding genes and may have phenotypic consequences. Twenty-nine putative gene conversions were identified in the chromosome substitution strains. We find that gene conversions are more likely to occur in regions where the homologous chromosomes are more similar. The raw genotyping data and genetic maps of these strain panels are available at http://churchill-lab.jax.org/website/MDA.

Introduction

The laboratory mouse is the most widely used mammalian model organism for biomedical research. Among the key advantages of mice are a well-annotated reference genome (CHINWALLA *et al.* 2002), over one hundred strain-specific genome sequences (KEANE *et al.* 2011), (MORGAN *et al.* 2016), (CC Genomes, Genetics 2017), and many genetic reference populations, including multi-parent strain panels (CONSORTIUM 2012) and outbred stocks (CHURCHILL *et al.* 2012), and strains carrying null alleles at most protein coding genes. There are dozens of readily available inbred strains that capture a wealth of genetic variants and display unique phenotypic characters (BECK *et al.* 2000), (YANG *et al.* 2011).

Genetic reference populations of mice include collections of strains that reassort a fixed set of genetic variants such as *chromosome substitution strain* (CSS) and *recombinant inbred strain* (RIS) panels. Chromosome substitution strains, also known as consomic strains, combine genomes of two founder inbred strains by substituting one chromosome pair from the *donor strain* into the genetic background of the *host strain* (NADEAU *et al.* 2012). The mouse genome is composed of 19 pairs of autosomal chromosomes, X and Y sex chromosomes, and a mitochondrial genome, thus a minimum of 22 strains could constitute a complete CSS panel. In some cases it has proven difficult to introgress a specific entire donor strain chromosome into the host background and the complete CSS panel may include partial chromosome substitutions and consists of more than 22 strains. RIS also combine genomes of two founder strains; they are derived from one or more generations of outcrossing followed by sibling mating to produce new inbred strains whose genomes are mosaics of the founder genomes (WILLIAMS *et al.* 2001). Both RIS and CSS panels have been successfully applied to the mapping of complex traits (BUCHNER and NADEAU 2015).

We have carried out high-density genotyping of three RIS panels C57BL/6J x DBA/2J (BXD), ILS/IbgTejJ x ISS/IbgTejJ (LXS), C57BL/6J x A/J (AXB/BXA) and three CSS panels C57BL/6J-Chr#<A/J> (B6.A), C57BL/6J-Chr#<PWD/Ph> (B6.PWD), C57BL/6J-Chr#<MSM/Ms> (B6.MSM) using the Affymetrix Mouse Diversity Array (MDA). The MDA includes approximately 623,000 probe sets that assay single nucleotide polymorphisms (SNPs) plus an additional 916,000 invariant genomic probes targeted to genetic deletions or duplications (YANG *et al.* 2009). These data add value to the strain panels by more precisely localizing the recombination breakpoints between founder strains. In addition they reveal some unexpected features in the genomes of individual strains.

Materials and Methods

Animals

We generated high-density genotype data for six mouse strain panels (Table 1): three panels of RIS and three panels of CSS. Mice for genotyping from five panels were available at the Jackson Laboratory (Bar Harbor, ME, USA) or from BXD colony at University of Tennessee Health Science Center (UTHSC); DNA samples from the sixth panel, B6.MSM CSS, were provided by T. Shiroishi (National Institute of Genetics, Japan). Unless stated otherwise, we genotyped one mouse per strain. Most strains are represented by a single male animal (255 males) but for four strains we genotyped an individual female (BXD14, BXD54, BXD59, BXD76). Samples were mainly from cases bred in 2008.

The AXB/BXA RIS panel (NESBITT and SKAMENE 1984) was derived from intercrosses of the C57BL/6J (B or B6) and A/J (A) strains. Note that hereafter the dam is denoted first and the sire last. Thus the difference between AXB and BXA strains is the direction of the intercross mating that generated (AxB)F1s or (BxA)F1s, respectively. We genotyped 25 strains: AXB strains 1, 2, 4-6, 8, 10, 12, 13, 15, 18, 23, 24; and BXA strains 1, 2, 4, 11-14, 16, 17, 24-26.

The LXS RIS panel (WILLIAMS *et al.* 2004) was generated at the Institute for Behavioral Genetics, Bolder, CO from founder strains, Inbred Long-Sleep (L or ILS) and Inbred Short-Sleep (S or ISS). These founder strains were in turn derived as selection lines from a cross population with eight founder strains (A, AKR, BALB/c, C3H/Crgl/2, C57BL/Crgl, DBA/2, IS/Bi and RIII). We genotyped 64 strains: LXS 3, 5, 7-9, 13, 14, 16, 19, 22-26, 28, 32, 34-36, 39, 41-43, 46, 48-52, 56, 60, 62, 64, 66, 70, 72, 73, 75, 76, 78, 80, 84, 86, 87, 89, 90, 92-94, 96-103, 107, 110, 112, 114, 115, 122, 123.

The BXD RIS panel was derived from founder strains C57BL/6J (B or B6) and DBA/2J (D or D2) inbred mice in three epochs: epoch I, strains 1-32 (TAYLOR *et al.* 1975); epoch II, 33-42 (TAYLOR *et al.* 1999), and the epoch III advanced RIS 43-102 (PEIRCE *et al.* 2004b). The latter were outcrossed for multiple generations before inbreeding. We genotyped 91 strains: BXD 1, 2, 5, 6, 8, 9, 11-16, 18-25, 27-36, 38-40, 42-45, 47-56, 59-71, 73-102 (note that the designation of several BXD strains have been modified as a result of the genotyping results described in the present study, and BXD103 is now known as BXD73b).

The B6.A CSS panel (NADEAU *et al.* 2000) consists of 22 strains derived from C57BL/6J (host) and A/J (donor) by J. Nadeau at Case Western Reserve University.

The panel includes 19 autosomes, X and Y chromosomes, and the mitochondrial genome.

The B6.PWD CSS panel (GREGOROVA *et al.* 2008) consists of 28 strains derived from C57BL/6J (host) and PWD/Ph (donor) by J. Forejt at the Institute of Molecular Genetics AS CR in Prague, Czech Republic, covering all chromosomes and the mitochondrial genome. To improve reproductive fitness, chromosomes 10, 11 and X were split between three strains each carrying either the proximal (p), middle (m), or distal (d) portion of the respective chromosome.

The B6.MSM CSS panel (TAKADA *et al.* 2008) consists of 29 strains derived from C57BL/6J (host) and MSM/Ms (donor) by T. Shiroishi at National Institute of Genetics in Mishima, Japan covering all chromosomes. Chromosomes 2, 6, 7, 12, 13, and X were split between two strains each carrying either the centromeric (C) or telomeric (T) portion of the respective chromosome.

Genotyping

DNA samples were prepared at the University of North Carolina according to the standard Affymetrix protocol and were hybridized on the Affymetrix Mouse Diversity Array (MDA) at the Jackson Laboratory as described previously in (YANG *et al.* 2009), (DIDION *et al.* 2012). The MDA probes (NCBI37/mm9) were mapped to genomic positions in GRCM38/mm10 assembly. CEL files and updated mapping information are available at ftp://ftp.jax.org/petrs/MDA/raw_data/. We used the R software package MouseDivGeno (DIDION *et al.* 2012) to extract intensities from CEL files, but for purposes of this study we developed a genotyping method that is based on the direct comparison of SNP probeset intensities between the sample and the founder strains of the corresponding panel. We selected the informative SNPs with intensity differences between founder strains for each panel (101,397 SNPs for AXB/BXA, 79,808 for LXS, 103,340 for BXD). Both selection of informative SNPs and SNP calls were probeset intensity based. For each strain and each SNP, the call can be either A (if the signal is close to the first founder), B (if the signal is close to the second founder), or N to represent “notA/notB”. We note that the N category includes both no-call and heterozygous genotypes and simply indicates that the intensity signal of the sample is far from both founder strains.

Founder Haplotype Blocks

In order to define the haplotype blocks of founder genotypes with allowance for errors in individual SNP level genotype calls, we applied the Viterbi algorithm to smooth the genotyping. We used software implemented in the Hidden Markov Model (HMM) R package (HIMMELMANN 2010). We call the Viterbi algorithm iteratively: at each iteration we re-estimated the HMM transition probabilities based on the Viterbi reconstruction of haplotype blocks. The iterations are repeated until we reach the convergence (JUANG and RABINER 1990).

Genetic maps computed from RIS panels consist of intervals assigned to one of the founders and gaps that delimit the interval within which the inferred recombination event(s) have occurred. We refer to the latter as “recombination intervals”.

For RIS panels we compared our maps to those available at http://www.genenetwork.org. GeneNetwork.org provides two genotype files for the BXDs—a "classic" set (pre-2017) of genotypes that have been used in most mapping studies since 2005 (SHIFMAN *et al.* 2006), and new consensus genotypes (2017) that include updated data for BXD43 through BXD220 that were collected November 2015 and processed using the GigaMUGA array (MORGAN *et al.* 2016). In the current study we have compared MDA genotypes to the classic genotypes used through the end of 2016.

Strain contamination

An RIS or CSS is considered to be contaminated if it carries a segment of genome that did not originate from one of the two founder strains. We developed an HMM to search for contamination. In contrast to our previous HMM analysis, here we select SNPs that were not informative (both founders have the same signal). In a contaminated region the signal of a given strain is expected to contain a higher proportion of SNPs that differ from both founder strains. To avoid only intervals covering three or more non-informative SNPs were reported.

Copy number variants

To determine if any of the RIS or CSS strains carried copy number variations (CNVs) that differed from the copy number in the founder strains, we applied the *simpleCNV* function of the MouseDivGeno package (DIDION *et al.* 2012). We accepted only those candidate CNV detections that had length >20kb and covered at least 10 IGP probes with *t*-statistic above 5 (p<1E-6).

Gene conversions

Gene conversions are short tracts (<1kb) of nonreciprocal transfer of genetic information between two homologs that occurs during meiosis. In the case of RIS, it is difficult to distinguish gene conversion events from short haplotype blocks that are due to closely spaced recombination events that occurred in different meiosis. Therefore we restricted our attention to the CSS panels. We searched for single or small groups of adjacent SNPs that derive from the host genotype but occur on the donor chromosomes. We examined individual SNP intensities to identify those that are clearly derived from the host strain and are present in a region of donor strain haplotype.

Sister strains

In a typical RIS panel the lineages that give rise to each RIS are independent and thus there should be no sharing of recombination events between strains. BXD strains from epoch III are an exception because they may share recombinations that arose in the outbreeding generations (PEIRCE *et al.* 2004a). Therefore, we excluded these strains from this analysis. We detected excess sharing of recombination junctions (Z-score>5.0) as an indicator that two strains are more similar than expected by chance.

Results

Global genotyping error - defined as a percentage of informative SNPs discordant with the haplotype assigment - is typically below 1%, but it is higher for haplotype blocks of *M. m. musculus* (PWD) and *M. m. molossinus* (MSM) origin than for *M. m. domesticus* blocks (B6, A, D2) (Suppl. Figure 1). This is likely to be caused by polymorphisms in or near the oligonucleotide probe sequence or its flanking restriction sites (DIDION *et al.* 2012). There are a few outlying strains with a higher error rate than other strains from the same panel (AXB1, BXD15, BXD25, BXD85, BXD65a (formerly known as BXD92), BXD93, B6.A#Chr7, B6.A#Chr10) likely due to low DNA quality or to processing of arrays.

Residual heterozygosity is present in some strains from each panel except for the AXB/BXA strains that appear to be fully inbred (Table 2). The detected heterozygous regions are an underestimate of percentage of segregating variation that is present in each strain because only a single animal per strain was genotyped. The presence of heterozygous strains in large RIS panels is not surprising. We estimated that in the absence of selection a RIS strain needs on average 24 generations of sib-mating to reach a heterozygosity rate below 1% and 36 generations to reach complete fixation. However, there is a significant variation in the number of generations required to achieve these landmarks (BROMAN 2005). For a panel of 22 strains (the size of a full CSS panel), 53 generations are required on average to achieve complete fixation for all its strains in the absence of selection.

***De novo* deletions and duplications:** We detected 64 *de novo* deletions and 14 *de novo* duplications, with lengths ranging from 21kb to 8.4Mb affecting 111 Ensembl genes (Suppl. Table 1). Table 2 summarizes frequency of strains with heterozygosity, deletions and duplications. We observe that longer time of inbreeding is associated with lower heterozygosity but more structural changes. This is seen most clearly by comparing different epochs of the BXD panel.

High-density genotyping identifies unexpected haplotype blocks in CSS panels

We observe 27 haplotype blocks from the host strain in the proximal or distal regions of the donor chromosome across the three CSS panels. These events are undesirable but not unexpected due to the distribution of markers used for CSS development (NADEAU *et al.* 2000). We also observe strains in which a host haplotype block occurs in the middle of an introgressed donor chromosome or a donor haplotype block occurs in a host chromosome. We observed seven such events distributed across all three CSS panels. See Table 3 for details.

High-density genotyping improves map accuracy in RIS panels

To validate our haplotype assignment and to estimate the level of improvement we compared our maps to the versions available at www.genenetwork.org (LXS, BXD) or provided by Institut de recherches cliniques de Montréal (AXB/BXA). There was a high concordance (99.8% LXS, 98.1% BXD, 99.5 ABX/BXA) between new and old maps for intervals that were in assigned to one of the founder in both maps. The new maps decreased the level of uncertainty measured as the sum of length of recombination intervals by 66% in the AXB/BXA panel, 41% in the BXD panel and 5% in the LXS panel. This improvement mirrors the increase in the number of informative markers: from 792 to 101,397 (AXB/BXA), from 3,796 to 103,341 (BXD), from 2,649 to 79,808 (LXS), respectively.

Strain contamination in the AXB/BXA panel

An unexpected observation in AXB/BXA RIS panel, was the presence of six intervals that are not derived from either A or B6 inbred strains. Three chromosomes of AXB1 (x, y and z), two chromosomes of AXB2 and one chromosome of BXA1 (x and y) are affected. Based on comparison to genotypes from a large panel of inbred strains (YANG *et al.* 2011) we conclude that the contamination derived from a strain that is closely related to DBA/2J.

Recombination rate

The distribution of the number of recombination events is similar across all panels (see Figure 1, Suppl. Table 2) with the exception of the advanced RIS BXD (epoch III) that has more recombination events per chromosome due to additional generations of outbreeding. The number of recombination events per strain ranges from 32 (BXD32) to 84 (BXA17) among the classical RIS and from 60 (BXD53) to 127 (BXD47) among the advanced BXD panel. These numbers of recombination events fall within the 95% prediction interval from simulations (using Python code from (WELSH and MCMILLAN 2012)).

Most recombination events in the RIS panels are unique but some recombination intervals overlap and could result from independent recurrent events or from shared ancestry between RIS during the inbreeding process. The most frequently shared recombination event occurs in 8 out of 25 samples of the AXB RIS panel (Chr10: 66,730,215-67,348,211). Moreover, in 7 out of 8 cases (p=0.07) the polarity of the event is in the same direction: from B6 segment (proximal - 66730214 bp) to A/J segment (67348212 bp - distal). Additional shared recombination intervals are listed in Suppl. Table 3 and the recombination frequency is visualized in Suppl. Figure 3. Higher recombination rates observed in the distal region of chromosomes are expected (LIU *et al.* 2014).

Sister strains

Sister strains are strains related by descent from incompletely inbred ancestors during the breeding process. They can be identified because they share a large number of recombination intervals with the same proximal to distal polarity of founder haplotypes. Not surprisingly, most of the sister strains are detected for the advanced BXD panel (6 pairs + 6 larger groups, totally comprising of 40 strains). However, two pairs of strains are present in the AXB and LXS panels, AXB6 - AXB12 and LXS94 - LXS107. These strains share more recombination intervals with the same founder strain polarity than expected by a chance (Figure 2).

The MDA array detects short gene conversions in CSS panels

We searched for putative gene conversions in the introgressed donor chromosomes of CSS panels. We identified small regions typically spanning just one informative SNP, that have genotypes consistent with the host strain instead of the donor strain

(Figure 3). In total, we identified 28 putative gene conversions: 17 in the B6.A CSS panel, 7 in the B6.PWD CSS panel and 4 in the B6.MSM CSS panel (Table 4).

Online access to genetics maps and MDA genotypes

For easy access, we provide a compilation of Mouse Diversity Array data, annotation and supporting software at http://churchill-lab.jax.org/website/MDA. Resources to support our analysis of RIS and CSS strains include an online viewer where maps can be viewed and downloaded either as a list of intervals or as CSV files ready to be imported to the R/qtl package (BROMAN and SEN 2009). Source code for the viewer is also available on Github, https://github.com/simecek/RIS-map-viewer. Researchers interested in comparing those reference populations to genotypes of other mouse strains processed on MDA arrays can used the MDA viewer. The entire database consisting of 1,902 MDA arrays is available for download as SQLite database or as individual CEL files ftp://ftp.jax.org/petrs/MDA/.

Discussion

We have characterized 180 RIS and 79 CSS strains from six popular and valuable resources and provided online access to these data. These panels were developed at different times and genotyped with lower density sets of markers. High-density genotyping with the number of informative SNPs, ranging between 79,000 and 257,000, provide maps with higher resolution. In this study we achieved a median spacing between informative markers 5.7 kb (AXB), 5.4 kb (BXD), 5.6 kb (LXS), 4.6 kb (B6.PWD) and 5.2 kb (B6.MSM), respectively. This enabled us to identify unusual features such as regions of residual heterozygosity, contamination by a non-founder strain and *de novo* structural variants. These genotyping arrays are part of 1902 samples processed on MDA platform that can be accessed from http://churchill-lab.jax.org/website/MDA.

Genetic reference panels are valuable, in part, because of the ability of generate animals with identical genomes in the number and timespan dictated by the researcher. Replication increases the accuracy of phenotype measurements (BELKNAP 1998) and allows for integration of data over space, time and environment. While it is convenient to think of all mice from an inbred strain as identical, we provide evidence that this view is not always warranted. Residual heterozygosity may be due to stochasticity in the inbreeding process or it may reflect biological constraints that prevent full inbreeding of a strain. Genetic drift operates in each of these populations and low-density genotyping in selected regions of the genome leaves room for undesired or unexpected surprises. In a typical CSS strain the average proportion of the donor genome present in other chromosomes is expected to be 0.2% (Nadeau 2000). Over our three CSS panels, the average length of unexpected genotype was 1.5 Mb. The length of intervals ranges (Table3) from less than 1 Mb (1 gene) to 20 Mb (138 genes).

For gene conversions, whole genome sequencing of CSS panels (and RIS) will likely reveal more examples and provide better estimates of converted regions and their length. However, our results suggest that gene conversions are more probable in regions where founders' genomes are very similar. We observe significantly more conversions on the B6.A panel that in the other two CSS panels (17 vs. 7 and 4, Fisher exact test, p=0.046) despite the fact that the number of informative markers is lower and therefore our ability to detect gene conversions reduced. Based on this result we hypothesize that gene conversions occur preferably in regions of low sequence diversity between homologous chromosomes. If that is true then they will have fewer genetic consequences due to lower chance to cause distinguishing polymorphism. Roughly, we estimate that 0.005% of the genome is affected by gene conversion (avg. # gene conversions / # informative SNPs = 28 / 3 / 200,000). The real number of gene conversions is likely to be higher because we were only able to identify gene conversions that overlap informative SNP probes in the array.

We found no evidence of allelic imbalance that has been observed in other species (TAUDT *et al.* 2016). Nor did we detect any epistatic selection between founder strains or alleles with different subspecies origin. This is in sharp contrast with mouse multiparent populations such as the Collaborative Cross and Diversity Outbred (CHESLER *et al.* 2016); (CC genomes 2017) and (SHORTER 2017). Due to limited number of strains in mouse RI panels, we may have missed small distortions.

We observed an inverse relationship between residual heterozygosity and drift (Table 2). For a given panel, even 20 generations of inbreeding is not enough to fix all heterozygous regions. On the other hand, populations kept for many generations will accumulate SNPs, small indels, and structural variants in their genomes (SIMECEK *et al.* 2015) (CC genomes 2017). Strategies to reduce drift in breeding colonies have been developed, including the embryo cryopreservation program at The Jackson Laboratory (TAFT *et al.* 2006). However, genetic drift can be also harnesed by geneticists to simplify and accelerate the identification of causal variants responsible for phenotypic differences between substrains (CC genomes 2017). These so call reduced complexity crosses are excellent examples of the potential benefits of genetic drift (KUMAR *et al.* 2013).

Acknowledgments

The work was supported by in part by NIH grants P50GM076468 (GAC), U19AI100625 (FPMV) and U01 AA016662 (RWW), a grant-in-aid for scientific research from Japan Society for the Promotion Science grant No. 17018033 (TS, TT), Canadian Institutes of Health Research MOP-93583 (CFD, MPSB) and by the Czech Science Foundation grant 16-01969S, LM2015040 and LQ1604 projects from the Ministry of Education, Youth and Sports of the Czech Republic (JF, PS). Genotyping of the BXD RIS was supported by the University of Tennessee Center for Integrative and Translational Science.

References:

BECK, J. A., S. LLOYD, M. HAFEZPARAST, M. LENNON-PIERCE, J. T. EPPIG *et al.*, 2000 Genealogies of mouse inbred strains. Nature genetics **24:** 23-25.

BELKNAP, J., 1998 Effect of within-strain sample size on QTL detection and mapping using recombinant inbred mouse strains. Behavior genetics **28:** 29-38.

BROMAN, K. W., 2005 The genomes of recombinant inbred lines. Genetics **169:** 1133-1146.

BROMAN, K. W., and S. SEN, 2009 *A Guide to QTL Mapping with R/qtl*. Springer.

BUCHNER, D. A., and J. H. NADEAU, 2015 Contrasting genetic architectures in different mouse reference populations used for studying complex traits. Genome research **25:** 775-791.

CHESLER, E. J., D. M. GATTI, A. P. MORGAN, M. STROBEL, L. TREPANIER *et al.*, 2016 Diversity Outbred Mice at 21: Maintaining Allelic Variation in the Face of Selection. G3: Genes| Genomes| Genetics **6:** 3893-3902.

CHINWALLA, A. T., L. L. COOK, K. D. DELEHAUNTY, G. A. FEWELL, L. A. FULTON *et al.*, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520-562.

CHURCHILL, G. A., D. M. GATTI, S. C. MUNGER and K. L. SVENSON, 2012 The diversity outbred mouse population. Mammalian genome **23:** 713-718.

CONSORTIUM, C. C., 2012 The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics **190:** 389-401.

DIDION, J. P., H. YANG, K. SHEPPARD, C. P. FU, L. MCMILLAN *et al.*, 2012 Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. BMC Genomics **13:** 34.

GREGOROVA, S., P. DIVINA, R. STORCHOVA, Z. TRACHTULEC, V. FOTOPULOSOVA *et al.*, 2008 Mouse consomic strains: exploiting genetic divergence between Mus m. musculus and Mus m. domesticus subspecies. Genome Res **18:** 509-515.

HIMMELMANN, L., 2010 HMM: HMM - Hidden Markov Models, pp. R package version 1.0.

JUANG, B.-H., and L. R. RABINER, 1990 The segmental K-means algorithm for estimating parameters of hidden Markov models. IEEE Transactions on Acoustics, Speech, and Signal Processing **38:** 1639-1641.

KEANE, T. M., L. GOODSTADT, P. DANECEK, M. A. WHITE, K. WONG *et al.*, 2011 Mouse genomic variation and its effect on phenotypes and gene regulation. Nature **477:** 289-294.

KUMAR, V., K. KIM, C. JOSEPH, S. KOURRICH, S.-H. YOO *et al.*, 2013 C57BL/6N mutation in cytoplasmic FMRP interacting protein 2 regulates cocaine response. Science **342:** 1508-1512.

LIU, J., H. SONG, D. LIU, T. ZUO, F. LU *et al.*, 2014 Extensive recombination due to heteroduplexes generates large amounts of artificial gene fragments during PCR. PloS one **9:** e106658.

MORGAN, A. P., J. P. DIDION, A. G. DORAN, J. M. HOLT, L. MCMILLAN *et al.*, 2016 Whole Genome Sequence of Two Wild-Derived Mus musculus domesticus Inbred Strains, LEWES/EiJ and ZALENDE/EiJ, with Different Diploid Numbers. G3: Genes| Genomes| Genetics **6:** 4211-4216.

NADEAU, J. H., J. FOREJT, T. TAKADA and T. SHIROISHI, 2012 Chromosome substitution strains: gene discovery, functional analysis, and systems studies. Mamm Genome **23:** 693-705.

NADEAU, J. H., J. B. SINGER, A. MATIN and E. S. LANDER, 2000 Analysing complex genetic traits with chromosome substitution strains. Nat Genet **24:** 221-225.

NESBITT, M. N., and E. SKAMENE, 1984 Recombinant inbred mouse strains derived from A/J and C57BL/6J: a tool for the study of genetic mechanisms in host resistance to infection and malignancy. J Leukoc Biol **36:** 357-364.

PEIRCE, J. L., L. LU, J. GU, L. M. SILVER and R. W. WILLIAMS, 2004a A new set of BXD recombinant inbred lines from advanced intercross populations in mice. BMC genetics **5:** 7.

PEIRCE, J. L., L. LU, J. GU, L. M. SILVER and R. W. WILLIAMS, 2004b A new set of BXD recombinant inbred lines from advanced intercross populations in mice. BMC Genet **5:** 7.

SHIFMAN, S., J. T. BELL, R. R. COPLEY, M. S. TAYLOR, R. W. WILLIAMS *et al.*, 2006 A high-resolution single nucleotide polymorphism genetic map of the mouse genome. PLoS Biol **4:** e395.

SHORTER, J., 2017 Designer protein disaggregases to counter neurodegenerative disease. Current Opinion in Genetics & Development **44:** 1-8.

SIMECEK, P., G. A. CHURCHILL, H. YANG, L. B. ROWE, L. HERBERG *et al.*, 2015 Genetic analysis of substrain divergence in non-obese diabetic (NOD) mice. G3: Genes| Genomes| Genetics **5:** 771-775.

TAFT, R. A., M. DAVISSON and M. V. WILES, 2006 Know thy mouse. TRENDS in Genetics **22:** 649-653.

TAKADA, T., A. MITA, A. MAENO, T. SAKAI, H. SHITARA *et al.*, 2008 Mouse inter-subspecific consomic strains for genetic dissection of quantitative complex traits. Genome Res **18:** 500-508.

TAUDT, A., M. COLOMÉ-TATCHÉ and F. JOHANNES, 2016 Genetic sources of population epigenomic variation. Nature Reviews Genetics.

TAYLOR, B. A., D. W. BAILEY, M. CHERRY, R. RIBLET and M. WEIGERT, 1975 Genes for immunoglobulin heavy chain and serum prealbumin protein are linked in mouse. Nature **256:** 644-646.

TAYLOR, B. A., C. WNEK, B. S. KOTLUS, N. ROEMER, T. MACTAGGART *et al.*, 1999 Genotyping new BXD recombinant inbred mouse strains and comparison of BXD and consensus maps. Mamm Genome **10:** 335-348.

WELSH, C. E., and L. MCMILLAN, 2012 Accelerating the inbreeding of multi-parental recombinant inbred lines generated by sibling matings. G3: Genes| Genomes| Genetics **2:** 191-198.

WILLIAMS, R. W., B. BENNETT, L. LU, J. GU, J. C. DEFRIES *et al.*, 2004 Genetic structure of the LXS panel of recombinant inbred mouse strains: a powerful resource for complex trait analysis. Mamm Genome **15:** 637-647.

WILLIAMS, R. W., J. GU, S. QI and L. LU, 2001 The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. Genome biology **2:** research0046. 0041.

YANG, H., Y. DING, L. N. HUTCHINS, J. SZATKIEWICZ, T. A. BELL *et al.*, 2009 A customized and versatile high-density genotyping array for the mouse. Nat Methods **6:** 663-666.

YANG, H., J. R. WANG, J. P. DIDION, R. J. BUUS, T. A. BELL *et al.*, 2011 Subspecific origin and haplotype diversity in the laboratory mouse. Nature genetics **43:** 648-655.

Tables and Figures:

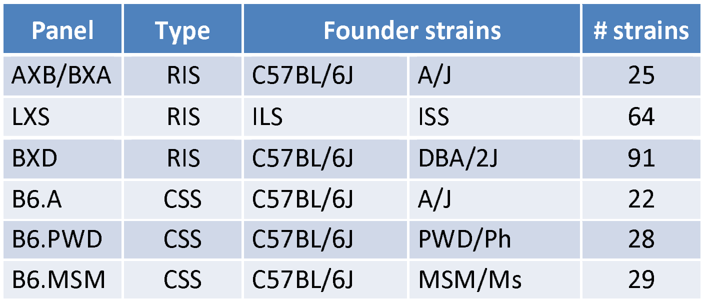
Figure 1: Number of founder haplotype blocks in RIS panels. The number of founder blocks for each strain is indicated as a point, with jitter for clarity. The boxplot indicates median and quartiles of each panel. Results for the BXD panel are broken down by three breeding epochs (I, II and III); the increased number of recombination event in epoch III refelects additional generation of outbreeding used in the derivation of these strains.

Figure 2: Sister strains in RIS panels. Side-by-side comparison of sister strains AXB6 vs AXB12 (red = B6, blue = A) (A) and LXS94 vs LXS107 (red = L, blue = S) (B) illustrates the extent of shared haplotype blocks.

Figure 3: Gene conversion in a CSS strain. Strain B6.PWD13 has an unexpected founder genotype at marker JAX00357227 marker (Chr 13: 47,505,217 bp). Average

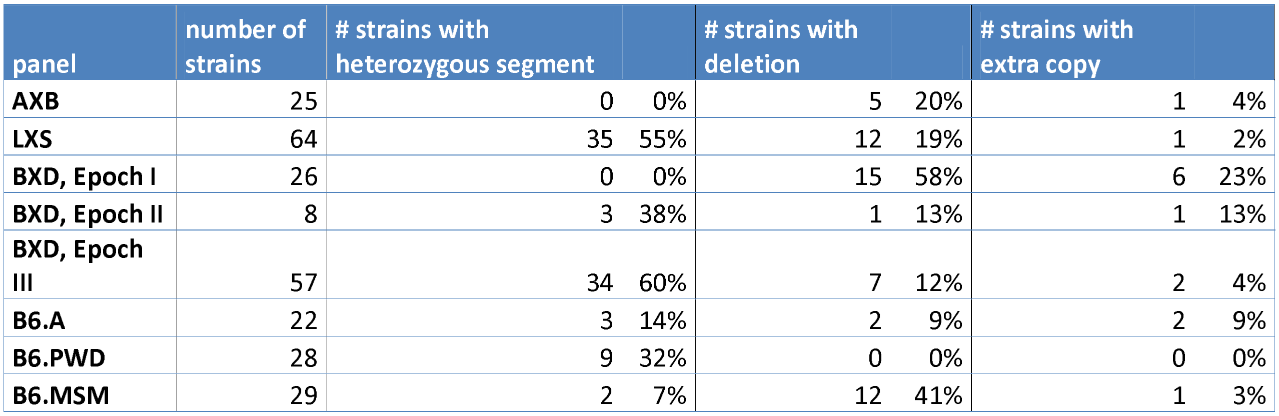
and contrast signal intensities are plotted for all B6.PWD strains. Numbers indicate the CSS strains by substituted chromosome with B6.PWD13 is highlighted by the red circle. Also indicated on the plot are founder strains B6, and PWD and their F1 hybrids. The B.PWD13 data should be similar to PWD but it is actually close to B6 indicating a putative gene conversion. Grey letters indicate genotype calls for 1902 additional samples in the MDA database (A = first parent / B = second parent / H = heterozygous / V = vino / N = no call).

Table 1: Overview of the six panels: a type, founder strains and a number of strains.



Panel	Type	Founder strains		# strains
AXB/BXA	RIS	C57BL/6J	A/J	25
LXS	RIS	ILS	ISS	64
BXD	RIS	C57BL/6J	DBA/2J	91
B6.A	CSS	C57BL/6J	A/J	22
B6.PWD	CSS	C57BL/6J	PWD/Ph	28
B6.MSM	CSS	C57BL/6J	MSM/Ms	29

Table 2: Residual heterozygosity and CNV (deletion / extra copy) in the six panels.



panel	number of strains	# strains with heterozygous segment		# strains with deletion		# strains with extra copy	
AXB	25	0	0%	5	20%	1	4%
LXS	64	35	55%	12	19%	1	2%
BXD, Epoch I	26	0	0%	15	58%	6	23%
BXD, Epoch II	8	3	38%	1	13%	1	13%
BXD, Epoch III	57	34	60%	7	12%	2	4%
B6.A	22	3	14%	2	9%	2	9%
B6.PWD	28	9	32%	0	0%	0	0%
B6.MSM	29	2	7%	12	41%	1	3%

Table 3: Unexpected haplotype blocks in all CSS panels.

Panel	Strain	Chr.	Start	End	Length	shoud be	actually is	# Ensembl genes
A.B6	C57BL/6J-Chr1A/J/NaJ	1	3211051	22830804	19619754	A	B6	117
A.B6	C57BL/6J-Chr1A/J/NaJ	1	192442075	195365691	2923617	A	B6	25
A.B6	C57BL/6J-Chr4A/J/NaJ	4	154799715	156166747	1367033	A or B6	Het	63
A.B6	C57BL/6J-Chr5A/J/NaJ	5	149410906	150567049	1156144	A or B6	Het	16
A.B6	**C57BL/6J-Chr8A/J/NaJ**	**11**	**36650633**	**42751289**	**6100657**	**B6**	**A**	28
A.B6	C57BL/6J-Chr10A/J/NaJ	10	127645772	129615258	1969487	A or B6	Het	102
A.B6	C57BL/6J-Chr16A/J/NaJ	16	93670025	98040454	4370430	A	B6	47
A.B6	C57BL/6J-Chr17A/J/NaJ	17	3071428	6154773	3083346	A	B6	25
PWD.B6	**C57BL/6J-Chr1PWD/ForeJ**	**3**	**123275916**	**143575204**	**20299289**	**B6**	**Het**	138
PWD.B6	C57BL/6J-Chr3PWD/ForeJ	3	24121111	24179212	58102	PWD	B6	1
PWD.B6	**C57BL/6J-Chr4PWD/ForeJ**	**5**	**148956085**	**151725288**	**2769204**	**B6**	**Het**	33
PWD.B6	C57BL/6J-Chr9PWD/ForeJ	9	123944659	124087880	143222	PWD	B6	3
PWD.B6	C57BL/6J-Chr10.1PWD/ForeJ	10	57607018	60613285	3006268	PWD or B6	Het	37
PWD.B6	C57BL/6J-Chr10.2PWD/ForeJ	10	45150578	51959138	6808561	PWD or B6	Het	21
PWD.B6	C57BL/6J-Chr10.2PWD/ForeJ	10	95379265	101638084	6258820	PWD or B6	Het	32
PWD.B6	C57BL/6J-Chr10.3PWD/ForeJ	10	73546548	74465198	918651	PWD or B6	Het	1
PWD.B6	C57BL/6J-Chr11.1PWD/ForeJ	11	3105931	3877120	771190	PWD or B6	Het	29
PWD.B6	C57BL/6J-Chr11.1PWD/ForeJ	11	79051423	79574667	523245	PWD or B6	Het	10
PWD.B6	C57BL/6J-Chr11.2PWD/ForeJ	11	35418368	43961733	8543366	PWD or B6	Het	53
PWD.B6	C57BL/6J-Chr11.3PWD/ForeJ	11	120588649	121967849	1379201	PWD	B6	60
PWD.B6	C57BL/6J-Chr12PWD/ForeJ	12	116831193	120014765	3183573	PWD	B6	16
PWD.B6	C57BL/6J-Chr19PWD/ForeJ	19	60070470	61261300	1190831	PWD	B6	20
PWD.B6	C57BL/6J-ChrX.3PWD/ForeJ	X	167416568	169593020	2176453	PWD	B6	16
MSM.B6	C57BL/6J-Chr4-MSM	4	24485868	24671707	185840	MSM	B6	3
MSM.B6	C57BL/6J-Chr6C-MSM	6	3180317	3410126	229810	MSM	B6	2
MSM.B6	C57BL/6J-Chr6T-MSM	6	147160651	149556829	2396179	MSM	B6	39
MSM.B6	C57BL/6J-Chr11-MSM	11	121160079	121967849	807771	MSM	B6	21
MSM.B6	**C57BL/6J-Chr12C-MSM**	**1**	**195151543**	**195285766**	**134224**	**B6**	**Het**	2
MSM.B6	**C57BL/6J-Chr13T-MSM**	**18**	**23532046**	**26528643**	**2996598**	**B6**	**Het**	24
MSM.B6	C57BL/6J-Chr14-MSM	14	122545401	124751019	2205619	MSM	B6	10
MSM.B6	C57BL/6J-Chr15-MSM	15	10241546	10362802	1212566	MSM	B6	46

			1	6				
MSM.B6	C57BL/6J-Chr16-MSM	16	95378122	98069653	2691532	MSM	B6	28
MSM.B6	C57BL/6J-Chr19-MSM	19	57068415	60681568	3613154	MSM	B6	27

Table 4: Short gene conversions in CSS panels.

Panel	snpId	chr	position	alleleA	alleleB	rsNumber	GeneSymbol	functionClass
B6.A	JAX00254769	1	72747910	C	T	rs50360495	N/A	Intergenic
B6.A	JAX00506852	2	1,49E+08	G	T	rs28225187	Napb	Intron
B6.A	JAX00517779	3	28468788	G	A	rs29689086	Tnik	Intron
B6.A	JAX00518655	3	31991151	G	A	rs49710262	N/A	Intergenic
B6.A	JAX00544220	4	7146585	C	T	rs27658062	N/A	Intergenic
B6.A	JAX00548886	4	41108534	C	T	rs27765251	N/A	Intergenic
B6.A	JAX00589927	5	1,01E+08	A	G	rs31987722	N/A	Intergenic
B6.A	JAX00630284	6	1,46E+08	C	A	rs30468531	Itpr2	Intron
B6.A	JAX00154063	7	89592185	G	A	rs51617084	N/A	Intergenic
B6.A	JAX00015582	10	20181498	G	C	rs29339980	Mtap7	Intron
B6.A	JAX00290764	10	62127533	C	T	rs46386144	N/A	Intergenic
B6.A	JAX00297554	10	1,03E+08	A	G	rs47130688	Lrriq1	Intron
B6.A	JAX00306860	11	30181044	G	A	rs26860826	Spnb2	Intron
B6.A	JAX00364408	13	81444533	G	A	rs29225071	Gpr98	Exon(Coding nonsynonymous)
B6.A	JAX00065772	15	1,03E+08	T	C	rs13482749	Map3k12	Exon(Coding synonymous)
B6.A	JAX00431551	17	11319650	G	A	rs33634737	Park2	Intron
B6.A	JAX00439159	17	44034125	A	G	rs33551899	Rcan2	Intron
B6.PWD	JAX00486683	2	########	A	C	rs28259595	5830434P21Rik	Intron
B6.PWD	JAX00507172	2	########	C	T	rs27373039	2310001A20Rik	Intron
B6.PWD	JAX00171651	9	########	C	T	rs30230810	Lman1l	Intron
B6.PWD	JAX00708417	9	########	C	A	rs36948070	Sacm1l	Intron
B6.PWD	JAX00357227	13	########	T	C	rs47221967	N/A	Intergenic
B6.PWD	JAX00072010	16	########	A	G	rs50630491	Cyyr1	Intron
B6.PWD	JAX00477099	19	########	T	C	rs31075313	Plce1	Intron
B6.MSM	JAX00250951	1	53216029	G	T	rs32733914	Pms1	Intron
B6.MSM	JAX00526581	3	72819995	G	A	rs37284921	N/A	Intergenic
B6.MSM	JAX00599346	5	1,4E+08	T	G	rs32296220	A930017N06Rik	Intron
B6.MSM	JAX00427113	16	87664971	C	A	rs47532274	N/A	Intergenic

Supplemental Figure 1: Error rate. Each strain is plotted by two dots of different colors (one dot = one founder strain). A dot represents a percentage of markers contradicting the estimated founder strain.

Supplemental Figure 2: Percentage of genome attributed to the first or the second RIS founder strain (red = B6 or ILS, blue = A/J or D2 or ISS, green = heterozygous)

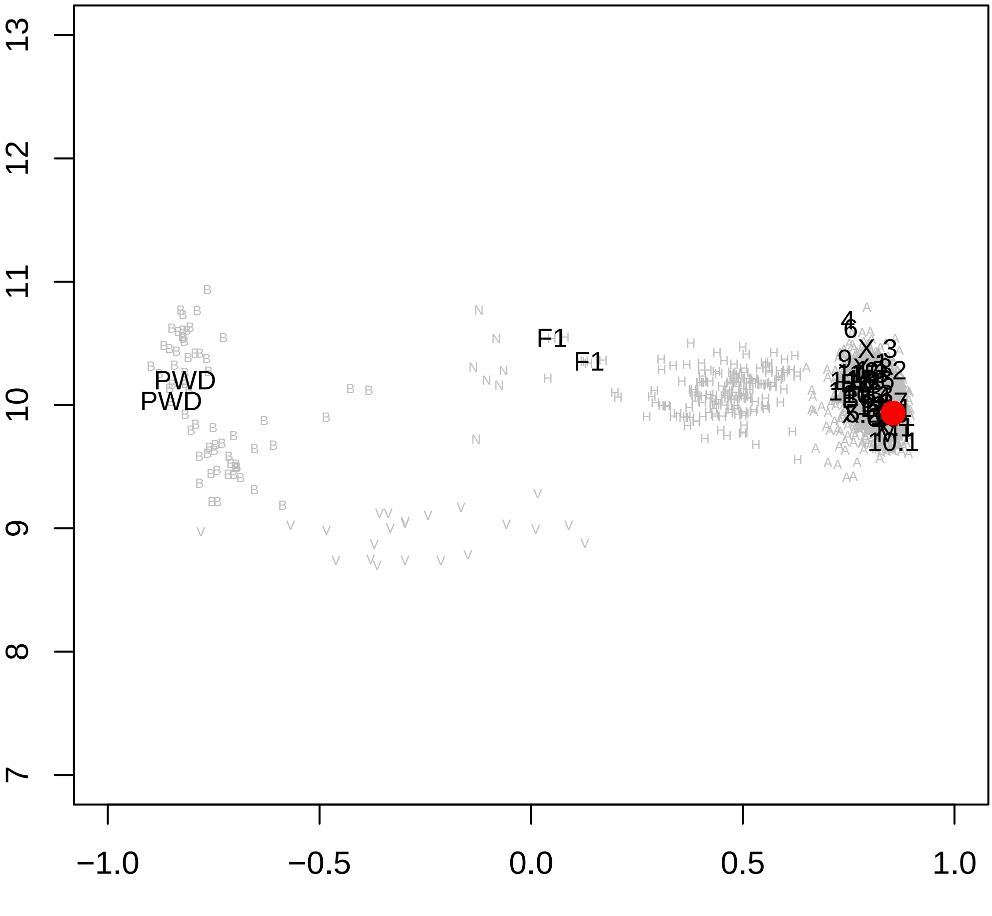
Supplemental Figure 3: Number of recombinations (smoothed by 10Mb window).

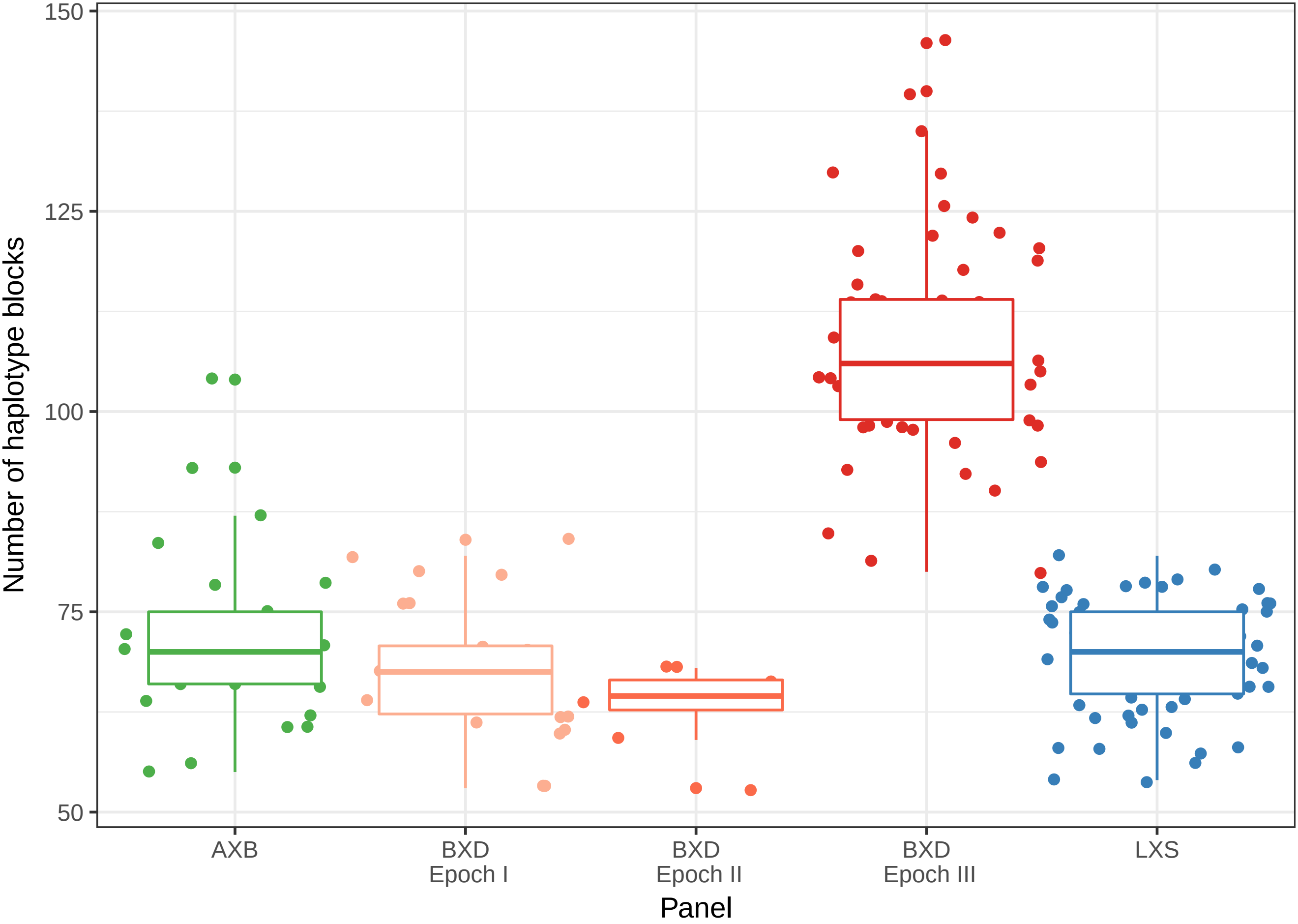
Supplemental Table 1: The list of RIS CNVs (deletion / extra copy).

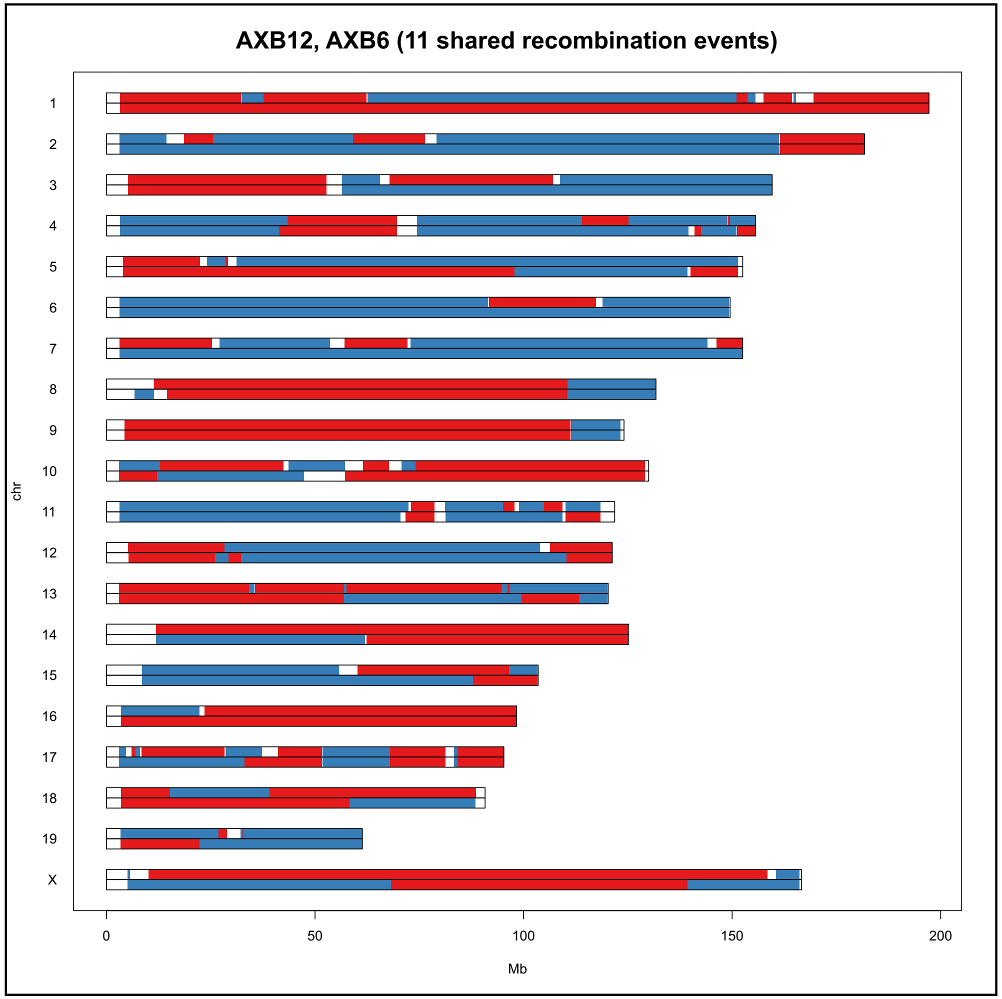
Supplemental Table 2: Number of haplotype blocks (first founder / second founder / heterozygous) and the total number of recombinations.

Supplemental Table 3: The list of all recombination intervals (and the frequency of recombination).

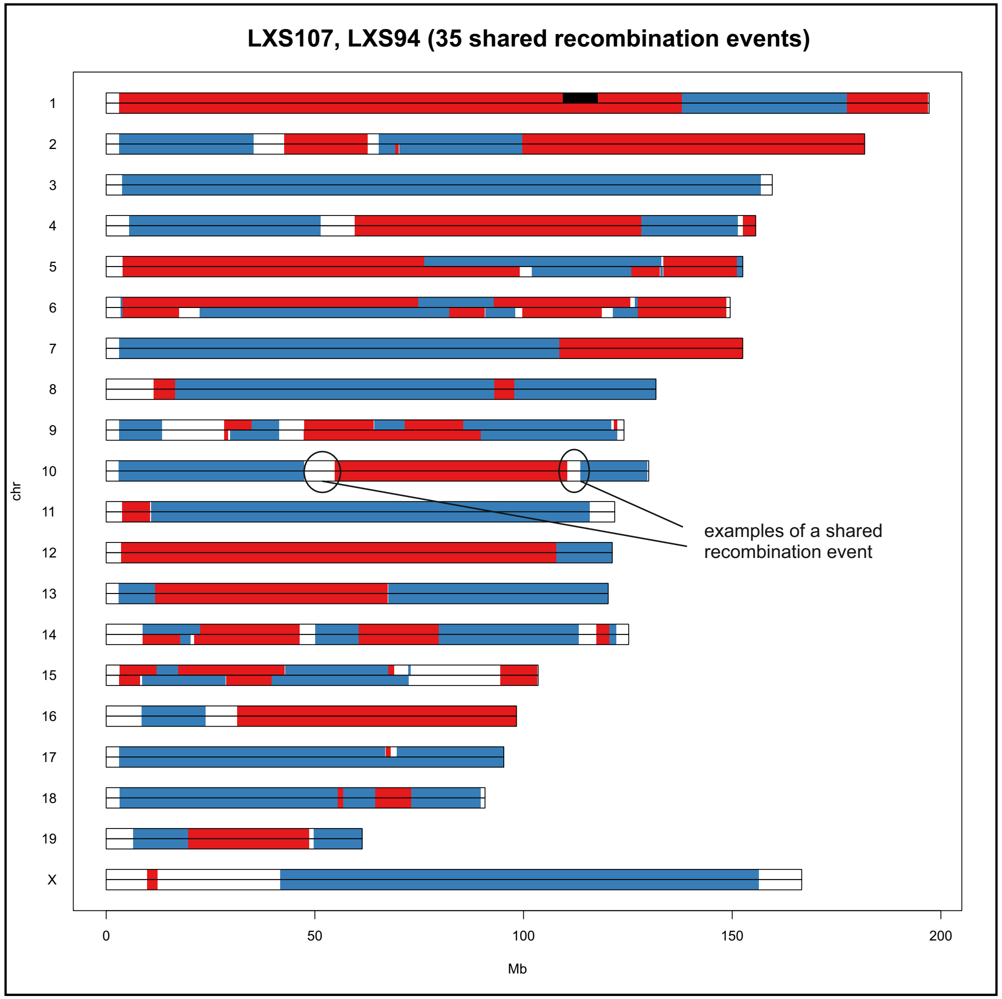
JAX00357227 (chr13: 47505217)




Number of haplotype blocks
150
125
100
75
50
AXB
BXD Epoch I
BXD Epoch II
BXD Epoch III
LXS
Panel



AXB12, AXB6 (11 shared recombination events)



LXS107, LXS94 (35 shared recombination events)

examples of a shared recombination event