Private and sub-family specific mutations of founder haplotypes in the BXD family reveal phenotypic consequences relevant to health and disease

Ashbrook, D.G.[1,*], Sasani T.[9], Maksimov, M.[5], Gunturkun, M.H.[2], Ma, N.[5], Villani, F.[1], Ren, Y.[3], Rothschild, D.[7,8], Chen, H.[2], Lu, L.[1], Colonna, V.[1,11], Dumont, B.[10], Harris, K.[9], Gymrek, M.[4,5,6], Pritchard, J.K.[7,8], Palmer, A.A.[3, 4], Williams, R.W.[1]

*Corresponding author

1. Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA
2. Department of Pharmacology, Addiction Science, and Toxicology, University of Tennessee Health Science Center, Memphis, TN, USA
3. Department of Psychiatry, University of California San Diego, La Jolla, CA 92093-0667, USA
4. Institute for Genomic Medicine, University of California San Diego, La Jolla, CA 92093-0667, USA
5. Department of Medicine, University of California San Diego, La Jolla, CA 92093-0667, USA
6. Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093-0667, USA
7. Department of Genetics, Stanford University, Stanford 94305 CA, USA
8. Department of Biology, Stanford University, Stanford, CA 94305, USA
9. University of Washington, Department of Genome Sciences, Seattle, WA
10. The Jackson Laboratory, 600 Main Street, Bar Harbor, ME, 04609, USA
11. Institute of Genetics and Biophysics, National Research Council, Naples, 80111, Italy

Abstract

The BXD recombinant inbred (RI) mouse strains are the largest and most deeply phenotyped inbred panel of vertebrate organisms. RIs allow phenotyping of isogenic individuals across virtually any environment or treatment. We performed whole genome sequencing and generated a compendium of SNPs, indels, short tandem repeats, and structural variants in these strains and used them to analyze phenomic data accumulated over the past 50 years.

We show that BXDs segregate >6 million variants with high minor allele which are dervied from the C57BL/6J and DBA/2J founders and use this dense variant set to define ‘infinite' marker maps and a novel family-level pangenome. We additionally characterize rates and spectra *de novo* variants which have accumulated over 20-200 generations of inbreeding, and have largely been ignored previously. Overall, the uniquely rich phenome when linked with WGS enables a new type of integrative modeling of genotype-to-phenotype relations.

.

Introduction

Recombinant inbred (RI) strains – inbred strains produced from crosses of two or more inbred parents – have been generated and used for a wide variety of applications in many species that tolerate inbreeding, including yeast[1], many plants[2–5], flies[6,7], nematodes[8], fish[9], rats and other rodents[10]. The BXD family is the largest and one of the oldest families of RI strains, started in the early 1970s by crossing female C57BL/6J (B6 or *B*) and male DBA/2J (D2 or *D*) inbred mice[11]. This family was produced by crossing the two inbred parents as if making a classical F2 intercross. The F2 progeny and all subsequent generations were then sibling mated to establish new "recombined" inbred lines. Each of the new derivative progeny strains has a unique but replicable genome that is a linear mosaic of ancestral haplotypes, inherited from *B6* and *D2* parents[12,13] (Figure 1A). The first set of ~26 BXDs were used mainly for mapping Mendelian loci[14,15], but as the family has grown—now to ~152 extant members—they have found new uses, including mapping complex traits (phenotype-to-genotype), reverse genetics using phenome-wide association (genotype-to-phenotype), understanding gene-by-environment and gene-by-sex interactions[11,16–18], and causal modeling[19].

Two factors set the BXD family apart from other vertebrate genetic reference panels. First, the number of family members is large enough for well powered and precise genetic analyses[11]. There are now 120 strains available directly from The Jackson Laboratory (Supplementary Table 1) with an additional 24 available from our laboratory. Second, the BXD have an extraordinarily rich multiomic phenome that has been accumulated over 50 years. Virtually all of these "polyphemone" data are available from a large and FAIR-compliant[20] web service (GeneNetwork.org). This dense and well-integrated phenome consists of over 10,000 classical phenotypes[21] and well over 100 molecular and expression QTL-type data sets, with the ability to replicate deeply and to extend across many treatments and environments[22,23].

Although BXD progeny are all nominally descended from the same two parents — C57BL/6J and DBA/2J— these parental strains will have accumulated their own new mutations over the 50 years since the creation of the first BXD strains. Furthermore, the BXDs themselves were produced in several epochs (Supplementary Table 1), some of which were produced from traditional F_2 intercrosses, and some which were produced using advanced-intercross progeny (Figure 1A and 1B). This has resulted in a more

complicated family genetic archetecture than is easily captured by genotyping arrays, but which is ideal for a family pangenome assembly.

Whole-genome sequencing has revolutionized all disciplines of biology. Since the first draft of the mouse genome in 2002[24,25], there have been important insights into the biology of the genome, the creation of new mouse models by genetic engineering[26], whole-genome screens, and the characterization of genomic diversity across mouse strains[27,28]. On a parallel course, stable mouse reference populations, such as the BXDs, have continued to be used for forward genetics, and with the sequencing of their parental strains, have also been used for systematic reverse genetics (also referred to as phenome-wide association studies)[18,29,30]. The present study is unique in that it is the first time a very large, fully isogenic family of strains have been deeply sequenced together with a large and open phenome—combining all the advantages of both forward and reverse genetics (although we note that inbred strains[28,31,32] and families have been sequenced previously)[33,34]. In the present study we have performed deep, linked read sequencing of all the BXD family, enabling us to characterize the full complement of genetic variation within each BXD strain and epoch, including not only SNPs and small indels, but short tandem repeats (STRs) and large structural variants (> 1000 bp). The result is an unprecedented, compendium of variants within a single family, including all autosomes and both sex chromosomes. We also provide new, updated, and curated genotypes for a total of 198 BXD family members—from BXD1 through to BXD220, including 44 strains that are now extinct but for which phenotype data have already been acquired (Supplementary Table S1), and whole-genome sequencing data for 152 of the extant strains. Deep DNA sequencing allows us to carry out genetic analyses that were previously not possible, such as identifying private variants that are *de novo* in single family members. A subset of variants are unique to BXD epochs, due to shared parentage. The variation in generations of full sib mating makes it possible to estimate mutation rates and spectra with high precision—so high that we can even map loci that control mutation rates and properties[35,36]. BXD progeny with extreme phenotypes can be caused by polygenic interactions, but in a number of cases are due to the impact of de novo variants—for example, blindness in the BXD24-Cep90 line[26–28]. What used to be a limitation—undefined mutations—has been converted into a major strength that bridges effectively between classic mutagenesis, genetic engineering, and classic quantitative gennetic mappoing approaches. The massive phenome makes it practical to identify not

only causal genes, but causal variants. This resource enables us to transition from genetic dissection to genetic prediction and synthesis of genotype-to-phenotype relations.

Results

Sequencing the BXD family

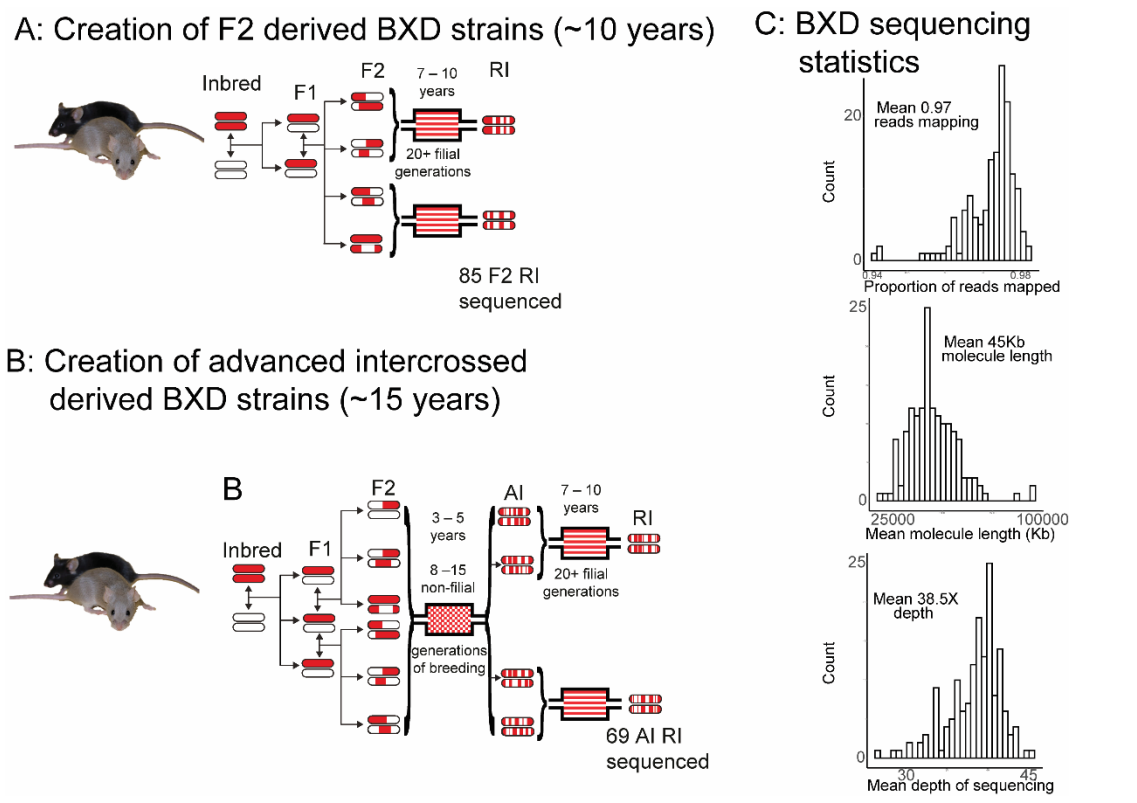


Figure 1: Overview of sequencing of the BXD family. Creation of the BXD family (**A**, **B**), and sequencing statistics (**C**). The epochs of the BXD family were created using either a standard F2 (**A**, n = 85) or by advanced intercrosses (**B**, n = 69) from the C57BL/6J and DBA/2J parents in several epochs, from the 1970s, and the oldest of these strains have been bred for ~200 generations (Supplementary Table 1). Red reprents regions of the genome inherited from the C57BL/6J, and white represents regions of the genome coming from the DBA/2J. Solid lines represent a single generation. Adapted from[11,17]. This design results in both strain-specific variants, and variants between epochs. To detect these we used link read sequencing from 10X Genomics ©, in which the genome is fragmented into ~40Kb fragments, and these fragments are barcoded before further fragmentation and short read sequencing. Sequencing metrics across strains are given (**C**).

We have carried out linked-read sequencing on 152 extant BXD strains, plus their C57BL/6J and DBA/2J parents (Figure 1). Sequencing produced a mean read depth of 38.6× and a mean molecule length of 44.5 kb (Figure 1C). Variants can be broadly

separated into segregating variants - that is variants that are inherited from the C57BL/6J and DBA/2J parents of the population and segregate in the BXD family, and *de novo* variants - variants that have occurred in a particular BXD strain or group of strains during production.

Segregating variants

SNPs and Indels

Using the GATK pipeline[37,38] and a set of true-positive variants in which the DBA/2J differ from the reference C57BL/6J, we identified 5,891,472 SNPs and 696,596 indels at a minor allele frequency (MAF) greater than 0.2, which we assumed to be inherited in the BXD family from the parental strains (Figure 2A, Supplementary figure 1A). We clearly saw genomic regions of high diversity between the two parental strains (and therefore in the BXD family), as well as regions that were close to identical-by-descent between the parental strains (and therefore with few segregating variants in the BXD family). The mean minor allele frequency for the autosomes was 0.44 (Supplementary figure 1B).

Short tandem repeats (STRs)

To examine STRs, we built a genome-wide reference set of STRs with repeat units of 2-20 bp in the GRCm38 (mm10) reference genome (Figure 2B)[39,40]. We then used GangSTR[39] to infer the repeat copy number at each STR in the reference in each BXD strain (Methods). Genotypes showed high call quality across repeat classes, with decreased quality at dinucleotide STRs (Supplementary Figure 2A) which are more prone to errors introduced by PCR[41] in the 10X workflow. The majority of STRs in our reference panel were mono-allelic (69.3%) in the BXD and thus were excluded from downstream analyses. Another 8% of loci were filtered out either because they overlapped known genomic duplicationss or had an insufficient call rate across the family.

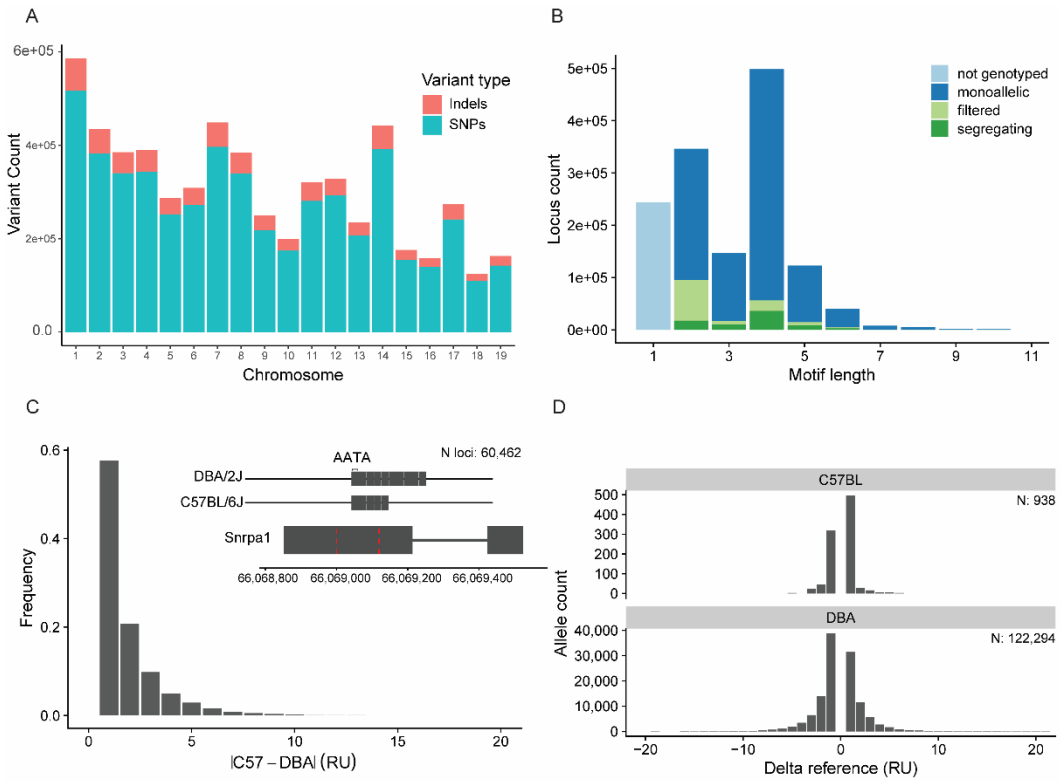


Figure 2: **SNPs, Indels and short tandem repeats (STRs) in the BXD family. (A)** Distribution of segregating (MAF > 0.2) SNPs and Indels across the autosomes of the BXD family **(B)** Description of STR reference panel. Bar height indicates abundance of STRs of a given motif length. Homopolymers (light blue) were omitted from genotyping and monoallelic loci (dark blue) were omitted from downstream analyses. Calls not passing quality control metrics (light green) were filtered out to generate a final call set of segregating STR loci (dark green). **(C)** Distribution of mutation size for STR loci relative to size of repeat in the reference for both parental strains—C57BL/6J and DBA/2J. **(D)** Distribution of differences in STR repeat counts between BXD founder strains and inset example locus with an AATA motif, overlapping an exonic region of a gene is shown as an example of a repeat where founder strains differ.

We used the resulting genotypes to characterize the extent of STR variation. As expected, D2 showed increased variation relative to the reference assembly (Figure 2C). Considering only homozygous calls, the parental strains varied at 60,462 (78.8%) of polymorphic STRs (Figure 2D) with a mean difference in repeat length of 1.9 units. The remaining polymorphic STRs not found in the parents were predicted to be *de novo*, and

are discussed below. While we found some larger repeat expansions and contractions of up to 10 repeat units, the majority had a difference of 1–2 units. Contractions relative to the reference allele were more abundant in D2 (55%/45%).

Genotypes for the majority (92–96%) of variable STRs analyzed in both parents and BXD progeny were inferred to be homozygous for one of the two parental alleles (Supplementary Figure 2C) and recapitulated the homozygous patchwork of inheritance blocks in agreement with the SNP results (Supplementary Figure 2E and 2D). Additionally, we observed an increase in heterozygous genotypes for family members from more recent epochs (Supplementary Figure 2C), consistent with incomplete inbreeding.

Large structural variants

We used SVJAM[42] to jointly genotype large structural variants across 152 BXDs. We analyzed 31,454 candidate SVs (deletions, inversions, and duplications) suggested by the LongRanger pipeline[43]. SVJAM uses image processing and convolutional neural networks to determine the type of SV. It employs a novel clustering algorithm for calling genotypes across all samples, and can distinguish heterozygous from homozygous variants.

We called 4,153 SVs (1,968 deletions, 1,106 duplications and 1,079 inversions). The number of SVs that were found in each chromosome is shown in Figure 3A. The minimum size of SV was 1,002 bp, the mean size was 78,832 bp, and the median was 33,272 bp. The distribution of sizes with respect to each SV type is presented in Figure 3C. In addition, we found SV hotspots including, a total of 272 SVs on Chr 13. Of these, 128 were located between 64–67 Mb, and 56 were located between 67–70 Mb . This is a region with a high density of annotated segmental duplications, likely facilitating high rates of homology based structural variation (Figure 3B).

Figure 3: Quantitative results of large structural variant detection. (A) Number of structural variants (SVs). Deletions (red) are the most common form of SV detected. Variants are spread across the genome, although there is significant variation between chromosomes. **(B)** Location of structural variants.There are SV hotspots throughout the genome. **(C)** Size distribution of large structural variants **(D)** Minor allele frequency of large structural variants

Infinite marker map

A smaller set of 5,271,335 autosomal and 37,830 sex chromosome SNPs called with high confidence were used as the basis of an 'infinite marker map' that effectively defines every recombination point to the interval between the two nearest informative variants. This allowed us to identify almost every haplotype structure for each genome, and to identify the first and last variant in each haplotype block. We combined this with previous

genotype-array data for extinct BXDs (Supplementary Figure 1C). The total number of recombination per BXD ranged from 26 to 125 (Supplementary Figure 1D). As expected, epochs derived from F2 crosses (epochs 1, 2, 4 and 6) have approximately half the average number of recombinations as epochs derived from advanced intercrosses (epochs 3 and 5).

The large numbers of recombinations in the BXDs, combined with the precise localization of each cross-over, increases the precision of QTL mapping. We reanalyzed the BXD phenome to demonstrate the advantages of these new genotype maps. For example, a phenotype that previously did not have a significant QTL (water intake of 13-week old females; GN BXD_12889), now has a significant and well defined QTL on Chr 9, encompassing only 10 coding genes (Supplementary Figure 3).

To estimate the overall precision of QTL localization, we used transcripts for genes on Chr 19, from four of the largest available BXD transcriptome studies of four different tissues, each of which contained at least 70 strains, encompassing 81 total distinct strains. Accuracy and precision were defined as the distance between the probe site and the peak QTL linkage value for all *cis*-eQTLs within 5Mb of the probe site. For these 1,525 transcripts, we plotted QTL accuracy as a function of the peak LOD score (Supplementary Figure 1E). Even for nominally significant loci (LOD 3.5), the mean distance is only 1.5 Mb, and accuacy improve to less than 1 Mb at LOD > 4. The median distance is even more precise —1 Mb for LOD > 3.5, and 0.5 Mb for a LOD > 5. This demonstrates that the distance between the peak of the QTL and the true causal variant will often be less than 1 Mb when using this new marker map.

De novo variants

Rates and spectra of new mutations - SNPs and indels

RI strains have always been used to for mapping common parental variants, whereas rare or spontaneous fixed variants have been largely ignored. However, each member of the family is expected to carry its own unique set of homozygoes mutations, which have arisen and been fixed over many generations of sib matings (Supplementary Table 1). To investigate rates and patterns of fixed *de novo* mutation among the BXD, we searched for all private homozygous singleton variants in each of family member. These singletons are by definition absence from both DBA/2J and C57BL/6J parents and all other BXD progeny.

We required these variants to be supported by at least 10 sequencing reads and to have a Phred-scaled genotype quality of at least 20. As expected, family members from earlier epochs have accumulated more homozygous singletons (Figure 4A). We additionally characterized the mutation spectrum of homozygous singletons, which describes the frequency of singletons corresponding to each of the six possible mutation types (C>A, A>G, etc.), in addition to C>T transitions at CG dinucleotides; overall, most singletons were C>T transitions (Figure 4B).

Some germline mutations might also arise earlier during the production of inbred strains and/or in particular founder individuals used for a particular epoch. Such mutations may be inherited by a subset of BXDs despite being absent from both the canonical C57BL/6J and DBA/2J parental strains. As each BXD epoch was initiated from a unique group of C57BL/6J and DBA/2J animals, we expected that these germline mutations would be shared more frequently by BXDs from the same epoch. On the other hand, *de novo* mutations shared by multiple BXDs from different epochs might represent recurrent

de novo mutations or common genotyping errors. For each pair of BXDs, we calculated the number of sites at which they share mutations with each other, divided by the combined number of sites at which those two BXDs share mutations with any BXDs (in effect, a measure of Jaccard similarity between the sets of shared mutations in each BXD). Overall, BXDs tended to share more mutations with mice from the same epoch, demonstrating that some germline mutations are epoch specific, and were inherited by multiple RI strains (Supplementary Figure 1F). Between epoch mutation sharing was observed between strains from epochs 4 and 5 (Supplementary Figure 1F), potentially reflecting that these epochs were initiated from the same cryopreserved founders (from the JAX genetic stability program)[44]. Those founders may possess unique mutations absent from the mouse reference sequence[45];

Rates and spectra of new mutations - STRs

STRs exhibit rapid mutation rates that are orders of magnitude higher than those of SNPs[46]. These mutations may occur at STRs that were fixed at a single allele in the founders or may occur at an STR that was polymorphic in the founders, leading to three or more alleles. Similar to other variant types, we expect recent *de novo* STR mutations

to be heterozygous, whereas mutations arising in ancestors to present-day strains are more likely to be homozygous as a result of inbreeding.

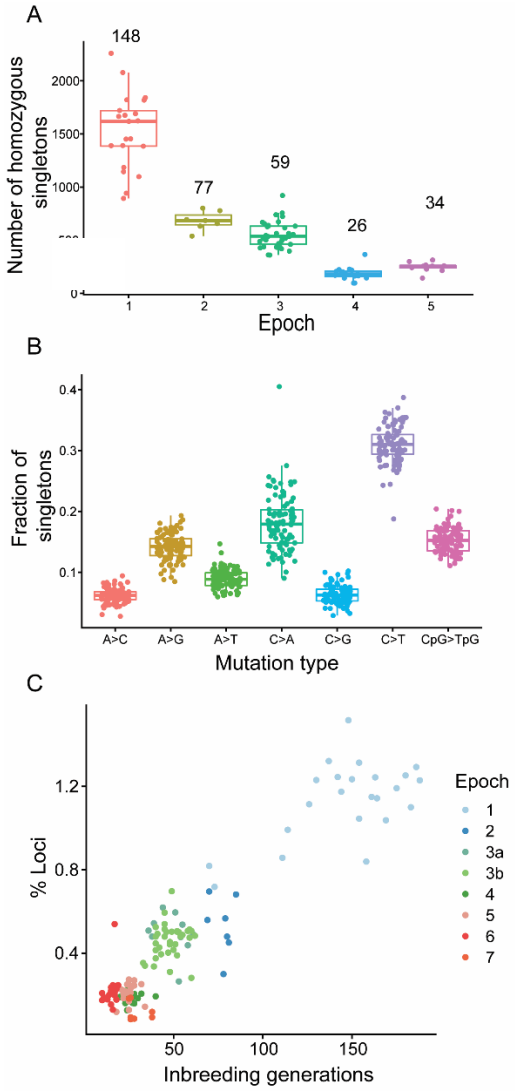


Figure 4: Germline mutation rates and spectra in the BXD family.(A) Counts of autosomal homozygous singleton SNPs in each BXD, stratified by epoch of origin. Average breeding generations are shown above each epoch. **(B)** Mutation spectrum of homozygous singleton SNPs across the BXDs, stratified by epoch of origin. Each homozygous singleton was classified according to its base substitution type, and C>T transitions were dichotomized by whether they occurred at CG dinucleotides. Mutations that were strand complements (e.g., C>T and G>A) were collapsed into a single mutation type. **(C)** Correlation between number of inbreeding generations and percentage of accumulated non-parental STRs

We identified new mutations at STRs by searching for loci where BXD genotypes did not match either of the founder genotypes. The percentage of new variant loci per strain is highly correlated with the number of inbreeding generations (Figure 4C). Earlier epochs contain more than 1.2% *de novo* loci per strain, while more recently derived strains contain fewer. Similarly to overall heterozygosity, the proportion of loci with one founder and one *de novo* allele increased in newer epochs, likely due to incomplete inbreeding (Supplementary Figure 2D). However, these private heterozygous variants may also be enriched for genotyping errors.

To get a more reliable estimate of the number of loci in BXD strains which are not found in the parental strains, we considered only STRs where at least one strain had a homozygous non-parental genotype resulting in the identification of 18,135 unique STR loci harboring mutations (Supplementary Figure 2D). As expected, the majority (54%) of non-parental mutations are singletons. The proportion of loci where we inferred a non-parental genotype for multiple strains drops off precipitously. As expected, many STR loci have three (the two parental alleles plus a mutation) or more alleles (Supplementary Figure 2B). We validated a subset of observed homozygous new mutations using fragment analysis across 16 strains at 27 unique STR loci. Fragment analysis matched GangSTR at 364/374 of calls (97%) available in both datasets. Across 39 total *de novo* mutations tested, 35 (90%) were confirmed by fragment analysis. The majority of discordant calls were at a single locus (Supplementary Figure 4) and differed by a single repeat unit.

Identification of a novel QTL regulating short interspersed element (SINE) transposition rate

To demonstrate our ability to find short interspersed elements (SINEs), we mapped B2 SINE counts using the newly generated marker map. We identified a significant QTL on Chr 5, peaking at 128.8 Mb, for the number of SINE elements. Although the QTL region was large (112.7 - 132 Mb) the peak was close to *Piwil1* (128.7 Mb), a gene that is known to promote the decay of a reporter mRNA containing a B1 or B2 SINE sequence[47].

Causal variant analysis

Some private variants are already known in the BXD, especially those which cause extreme phenotypes. One example of this is the BXD24/TyJ-Cep290^{rd16}/J strain. In 2004

a spontaneous mutation occurred in the BXD24/TyJ strain, causing a blindness phenotype[48–50]. Fortunately, frozen embryos from an earlier generation of the BXD24/TyJ strain were cryopreserved in 1988, allowing identification of an in-frame deletion in the causal gene, *Cep290*. The post-2004 strain was renamed BXD24/TyJ-Cep290^{rd16}/J. Our study sequenced both of these strains, allowing us to confirm the variant and its position.

Whole-genome sequence enables rapid identification of causal variants using either a forward genetics approach (seeing variation in a phenotype, and finding the causal variant, as with BXD24/TyJ-Cep290^{rd16}/J), or a reverse genetics approach (finding a genetic variant, and identifying phenotypes linked to it). Using variants identified in this paper, we provide examples of both approaches.

For example, we used a large liver proteome dataset, and carried out Rosner Outlier Tests to identify outlier strains for each peptide. Among the results, we saw that BXD63 is an outlier for expression of three peptides (Q8BJY1_LEAPLEELR_2, Q8BJY1_VFTAIDQPWAQR_2, Q8BJY1_ELTGEDVLVR_2), all of which are part of the protein PSMD5 (proteasome 26S subunit, non-ATPase 5). We also saw that BXD63 had low expression of *Psmd5* in several liver transcriptome datasets. Examining our singleton variants in BXD63, we find a single variant, an insertion of a C (A > AC) at chr2:34860673, which Variant Effect Predictor (VEP) annotates as a regulatory region variant. We hypothesize that this insertion is the cause of low *Psmd5* expression and low levels of PSMD5-derived peptides observed in BXD63.

BXD29, for which we sequenced two ‘sister’ strains, is an excellent example of how our new sequence data can solve old questions. It has previously been discovered that the BXD29-*Tlr4*$^{lps\text{-}2J}$/J mouse strain has a highly penetrant bilateral nodular subcortical heterotopia and partial callosal agenesis, and that it is not caused by the *Tlr4* variant[51,52]. In our sequencing of the BXD, we have identified only 975 variants that are present in the BXD29-*Tlr4*$^{lps\text{-}2J}$/J mouse strain that are not shared by the BXD29/TyJ strain - a strain derived from cryopreserved bankstock of F39 from 1979. From breeding records, we know that a spontaneous mutation arose in the BXD29-*Tlr4*$^{lps\text{-}2J}$/J between 1998 and 2004[51,52]. Given the large effect of the variant, we hypthesize that it is protein-coding, leaving only 30 of the 975 variants as candidates. These include a large duplication (chr11:74,689,025-75,132,234) affecting seven genes (*Mettl16*, *Mnt*, *Pafah1b1*, *Sgsm2*, *Smg6*, *Srr*, *Tsr1*). Importantly, in the human genome, duplication of this region, especially *Pafah1b1,* causes

cortical defects[53,54], including subcortical heterotopia[55,56]. A F_2 cross between the BXD29/TyJ and BXD29-*Tlr4*$^{lps\text{-}2J}$/J (sometimes termed a reduced complexity cross[57]) could be used to formally confirm this.

Structural Variation on the BXD Y Chromosome

The house mouse Y chromosome is dominated by several families of spermatid-expressed genes that are interspersed along the chromosome and present at upward of 100 copies. As a consequence of this genomic architecture, the Y chromosome is presumed to be vulnerable to homology-mediated expansions and contractions, although the paucity of high-quality Chr Y data from cohorts of related animals has made this phenomenon difficult to assess. Most BXD strains are derived from an initial cross between a B6 mother and a D2 father and should, therefore, harbor a D2-like Y chromosome, however 9 strains have a B6 father (see Supplementary Table 1). We took advantage of the 148 high-quality *de novo* assembled Chr Y genomes in the BXD to monitor Chr Y structural evolution .

We first sought to identify large-scale architectural features distinguishing the D2 and B6 Y chromosomes. We profiled the depth of coverage across Chr Y and observed a striking reduction in the density of D2 reads mapping to Chr Y relative to B6 (Supplementary Figure 5). This drop in read coverage is not uniformly distributed across the D2 Chr Y, but rather focused at discrete loci in the ampliconic portion of the chromosome (Supplementary Figure 5). This patterning is consistent with the presence of multiple large deletions that distinguish the D2 Chr Y from that of B6 (or, alternatively, multiple large insertions on B6 Chr Y), rather than a uniform drop out of reads due to D2 sequence divergence from the B6-based reference assembly. Indeed, copy number at two of the most highly ampliconic Chr Y gene families in mouse – *Sly* and *Ssty1/2* – is markedly lower in D2 than B6 (*Sly*: 124 copies in B6 versus 85 copies in D2; *Ssty1/2*: 437 copies in B6 versus 251 copies in D2). Based on the relative ratios of Chr Y reads to autosomal reads in these two genomes, we estimate the total length of the D2 Chr Y to be ~33.88 Mb shorter (36.9%) than the B6 Chr Y.

We extended this analysis to examine the scope of chromosome Y size variation in the BXD. Comparing the relative ratio of mapped Chr Y reads to autosomal reads in each sequenced strain and standardizing by the known length of the B6 Y chromosome

exposes remarkable variation in Chr Y size across the BXD family (Figure 5). Overall, approximately 77% of strains harbor Y chromosomes >5Mb longer than the parental D2 chromosome, with the Chr Y in nine strains even exceeding the B6 Chr Y length. In contrast, a few strains carry Y chromosomes smaller than the D2 parent (n = 6; Figure 5), suggesting a general bias toward Chr Y expansion, rather than contraction. At the extremes, the BXD157 Chr Y measures just 36.6Mb, whereas the Chr Y in BXD9 extends for 111.3Mb. Examination of depth of coverage across the Y chromosomes of BXD157 and BXD24, a second small Chr Y strain, (Supplementary Figure 5) reveals a general reduction of reads mapping to Chr Y, potentially indicative of a mosaic Chr Y loss in these sequenced samples, rather than bonafide changes in chromosome length. However, in other strains, the estimated small Chr Y size is clearly due to localized deletions: BXD13 harbors a ~38Mb deletion that extends from approximately chrY:28.65-67.8Mb and BXD27 exhibits multiple large deletions (Supplementary Figure 5). The large Chr Y size in BXD9 tracks to numerous large duplications across independent Chr Y segments in this strain (Supplementary Figure 5). On average, strains from epochs 1-3 harbor larger Y chromosomes than strains from later epochs (mean epochs 1-3: 77.93Mb, mean epochs 4-6: 63.85Mb; Mann-Whitney U-test: $P = 7.21 \times 10^{-18}$). Overall, these findings seem to point to a remarkable loss of Chr Y DNA during the maintenance of the inbred D2 strain, underscore the instability of this chromosome. Further studies will be needed to assess the functional impacts of this Chr Y size variation.

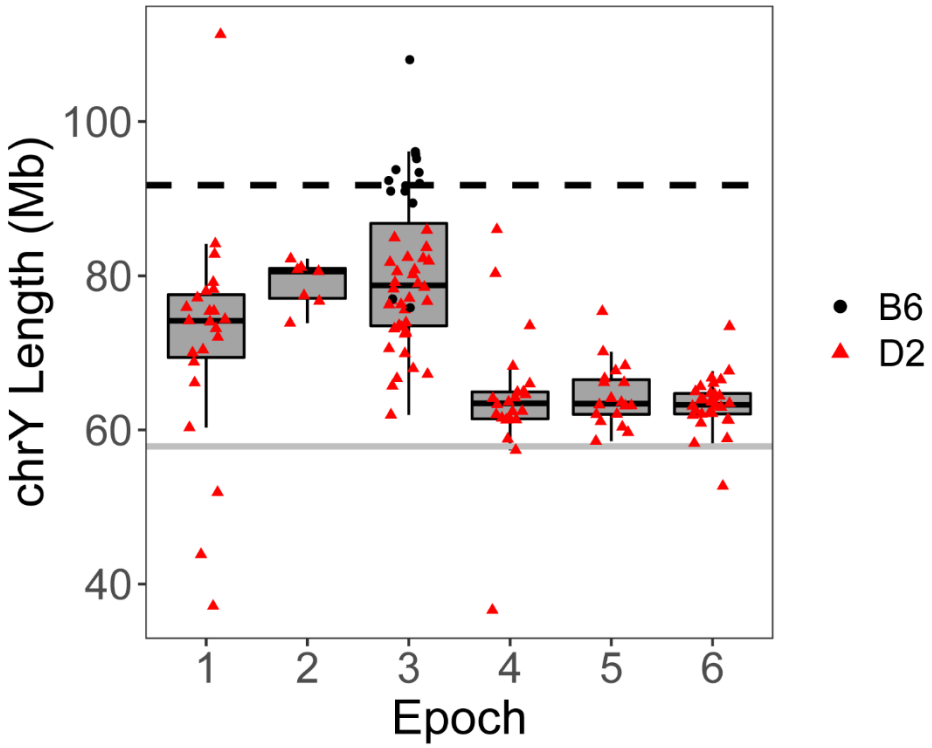


Figure 5. Distribution of Y chromosome length across the 6 BXD epochs. Y chromosome lengths were estimated by comparing the GC-corrected depth of coverage across Chr Y to the autosomes for each strain, followed by standardization to the known, fixed length of the B6 Chr Y. The dashed horizontal black line corresponds to the B6 Chr Y length. The solid gray horizontal line denotes the D2 Chr Y length. Strains carrying a DBA/2J derived Chr Y are shown as red triangles, strains carrying the C57BL/6J derived Chr Y are shown as black circles.

Pangenomic analysis identifies complex variants and is informative about strain-specific haplotypes

All the above work has used the classic approach of aligning to a reference genome. However, pangeomes[58–60] have key advantages, allowing all known variants to be included as part of the alignment process. As such, we built the BXD pangenome of Chr 19. This is expected to improve mapping of sequencing data (e.g. whole genome sequencing, RNA-seq and methylation-seq), and allow us to discover new variants. The BXD pangenome for Chr 19 consisted of 5.5M nodes and 8.6M edges (total length

264,380,676 bp with 82,695 paths, Figure 6A), starting from haploid assemblies of 10X reads. Four strains were excluded for poor assembly quality, leaving 148 strains in the pangenome.

We first focused on genetic variants up to 50bp (microvariants; SNPs and small indels). In DBA/2J we identified 182,643 variable sites. Of these, 178,557 are simple variants (SNPs and indels), and 4,086 are complex variants (allelic variants that overlap but do not cover the same range; Supplementary Table 2), many of which were not detected by GATK in our reference based variant calling (Figure 6B). This shows that the pangenome enables calling of complex variants not seen by traditional genomics methods such as clumped multiple nucleotide polymorphisms (MNPs) and MNP/INDEL (Supplementary Table 2). The Ts/Tv ratio in the pangenomic calls is 2.08, which is slightly lower than figures from GATK (2.15 and 2.09 for 10X and PacBio data, respectively), reflecting the enrichment of complex variant calls in the pangenomic set. Within masked regions (i.e. regions that do not contain SINE, ALUs, LINE, LTR, and other DNA repeats), precision and sensitivity of vg[61] calls are 90% and 85-86%, respectively, when compared to GATK. For SNPs these figures rise to 93-94% and 84-85% (Figure 6C, Supplementary Table 3), matching precision obtained from the comparison of the two truth sets (Supplementary Table 3).

Considering variants >50bp, we identified 61,381 variable sites. One example of these is a 2kb insertion (Figure 6D) in the second intron of the E3 ubiquitin ligase *Zfp91* (Figure 6E). This insertion was described in the DBA/2J strain[28] and is found in 48% of our BXD mice in complete linkage with two other insertions of 4 bp and 135 bp on a haplotype spanning 2,789 base pairs (Figure 6F). Using the presence or absence of the SV as a phenotype in GeneNetwork, we correlated this against the BXD phenome. The second most significant correlation was with 'Thymic T-cell proliferative unresponsiveness (anergy) to anti-CD3-induced proliferation' (p = 1.095e-04, r = 0.719). The same association between *Zfp91* and T-cell proliferative has recently been made by others in gene knockout mice.[62] The *Zfp91* gene facilitates TCR-dependent autophagosome formation to sustain T reg cell metabolic programming and functional integrity. *Zfp91* deficiency attenuates the activation of autophagy and associated downstream pathways and impairs T reg cell homeostasis and function, rendering mice sensitive to colonic inflammation and inflammation-driven colon carcinogenesis[63]. Therefore, there is strong evidence that segregating variants in *Zfp91* in the BXD cause alterations in T-cell

proliferation, and we expect this to result in clinically relevant phenotypes, e.g. relating to cancer and the immune system.

Overall, while the ideal data for an effective pangenome construction is long-reads, we conclude that the pangenome produced by short linked-reads can be informative.

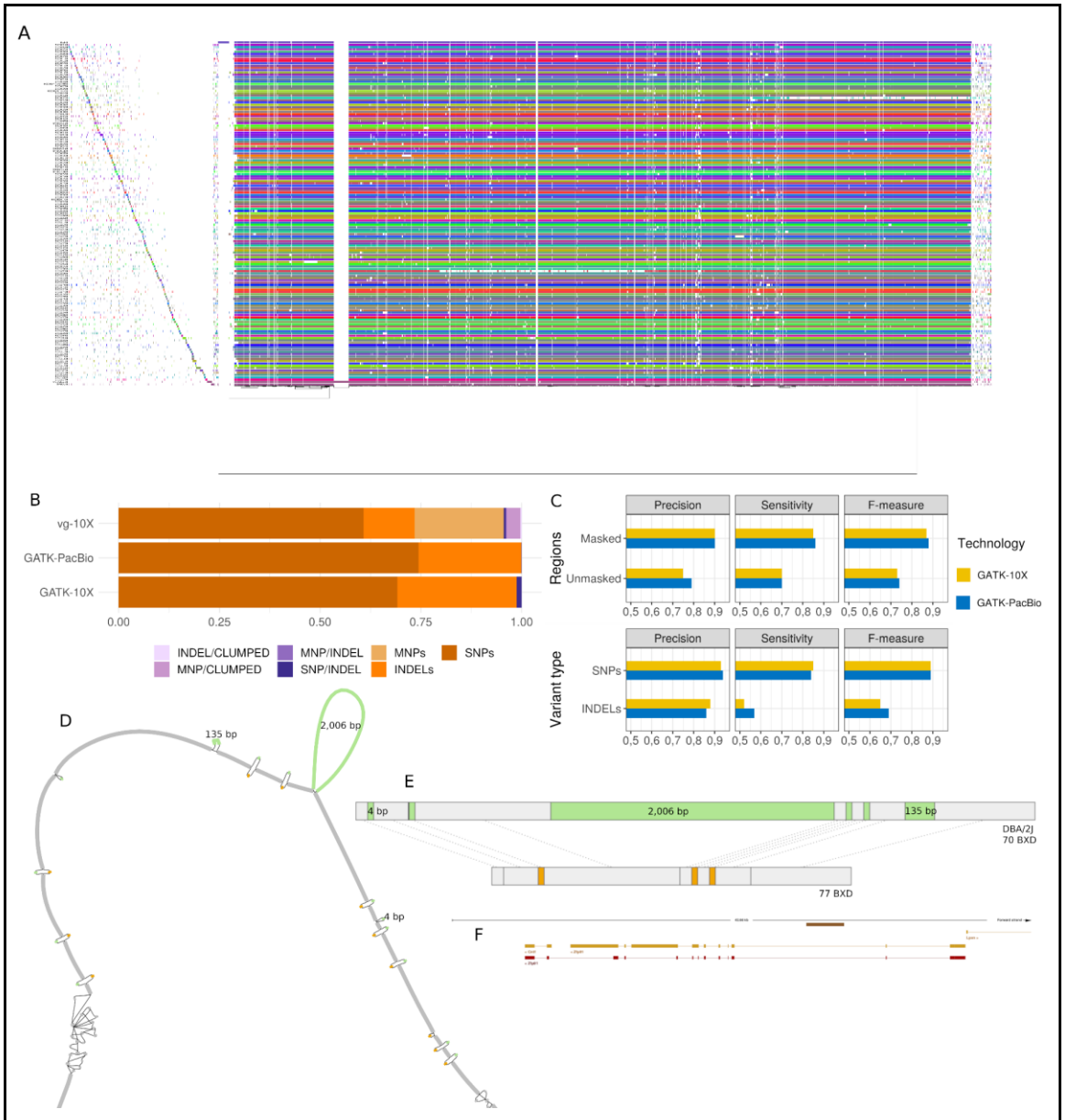


Figure 6: Pangenome of the BXD family. (A) *odgi-viz* linear visualization of the pangenome of Chr 19. Each line represents a haplotype. Line interruptions (white) are insertions in one or more strains, therefore deletions in the others (vertical white stripes). The left side is the centromere, the right side is the telomere. In these two regions sequences are fragmented. **(B) Types of microvariants** (up to 50bp) at variable sites in the pangenome call set (vg-10X), and the two truth sets (GATK-PacBio and GATK-10X). Simple variants are in shades of orange,

complex in shades of purple. *vg* call set includes the higher fraction of complex variants**. (C) Evaluation of *vg* call set in DBA/2J** using two truth sets sequenced with different technologies and called with GATK. Evaluation by variant type is within masked regions. **(D) Extract of the pangenome from the *Zfp91* gene** showing a 2,006 bp insertion found in DBA/2J and 48% of the BXD strains (green nodes in the graph). The insertion is in complete linkage with two other insertions of 4 bp and 135 bp in a region spanning 2.8 kbp. **(E) Strain-specific haplotypes** (gray segments are not in scale) **(F) Pangenomic extract in the gene context** is represented here by the brown segment above the second intron of the gene. The region on display corresponds to Ensembl *M.musculus* version 102.38 (GRCm38.p6) chr19:12,758,303-12,800,960.

Discussion

We report the sequencing of the entire BXD family, bringing deep, whole-genome, linked-read sequencing to this widely used and deeply phenotyped resource. We have identified SNPs, small indels, large structural variants, short tandem repeats and SINEs.

Using the known family structure of the BXD and the deep phenome that has been collected, we are able to link these variants to phenotypes, both at the family level and at the individual level. We show how these can be used to find new, molecular phenotypes, and to improve our ability to find causal genes for classical phenotypes.

The BXDs are well suited to studying gene-by-environmental interactions (GXE) and for experimental precision medicine[11]. For almost 50 years, they have been used to study the genetic and environmental factors that underlie a diverse collection of phenotypes, including environmental toxicant exposures, alcohol and drugs of abuse[64–68], infectious agents[69–73], diets[74–79], and stressors[80,81]. Beyond this, there are also extensive -omics data available for many BXD family members, including over 100 transcriptome datasets (e.g, [82,83]), as well as more recent miRNA[84,85], proteome[86–88], metabolome[83,88,89], epigenome[16,90], and metagenome[91,92] profiles. As each of these new data sets is added, it can be integrated with previous data, thereby multiplicatively increasing the usefulness of the whole phenome. We can easily identify strains that are outliers for any particular trait or molecular phenotype, and immediately have a short-list of candidate variants.

Although the BXD have been well characterized by array-based methods, allowing a long history of QTL mapping, this is the first time that the full breadth of variants have been available for the family. SINEs[93], CNVs[94], and large deletions[18] have been found in

the family previously, but these required extensive work to get from locus to variant. With the whole catalogue of variants available this is now only an afternoons' work.

Using deep genome sequencing data, we have also been able to investigate variation in mutation rates and spectra across the BXD family[25]. These called variants can become molecular phenotypes, which themselves can be mapped to causal loci. For example, we are able to identify genes causing specific mutation signatures[35,36], and a loci influencing the number of SINE in the genome. We give a brief overview of some of these analyses here, but they are an interesting and important outcome in and of themselves, and will provide fuel for future research.

The BXD are an enduring resource: phenotypes collected in the 1970s can be mapped using genotypes identified by our study, leading to the identification of novel QTLs. By extension, phenotypes collected in 2022 will still be informative in 2072. This is made possible not only by the use of well-characterized inbred strains, but by the dedicated community who have donated their data to open access resources, particularly GeneNetwork[22,23,95–97]. In this way, phenotypes collected across decades, continents and environments can be coherently coanalyzed, providing new insight from legacy data[95].

Furthermore, by crossing two BXD strains any of 22,350 isogenic F1 hybrids can be generated – a massive diallel cross (DAX)[11,98]. Thus, there are well over 22,000 reproducible "clones" of F1 hybrids that can be generated from the current BXD families, each of which carries one chromosome from each of its BXD parents, and is therefore fully '*in silico*' sequenced in advance. This will be a huge resource for the understanding of indirect genetic effects, gene-by-environment interactions, and epistasis. These can also be crossed to any other inbred strain, such as the Collaborative Cross population[33,99–103], or genetically engineered mice[104–106], expanding the amount of variation, and allowing for identification of modifier alleles. This has been excellently demonstrated with the AD-BXD[107–110], and others[111,112].

Given the deep and well recorded history, with over 180 generations of inbreeding in some strains, we have a unique resource. This project revitalizes this 50-year-old family, allowing many new analyses beyond the sample given here, and exponentially expands the utility of over data collected over many decades. This data not only allows the identification and genetic mapping of 'molecular phenotypes' (e.g. mutation spectra), but

is fully interactable with the whole BXD phenome, and is available for all future users of the BXD family.

Acknowledgements

We thank Dr. Benjamin A. Taylor for initiating the BXD and for continued support and encouragement. We thank Drs. Gerald McClearn and Lisa Tarantino for 60 eighth-generation (G8) AI progeny from Pennsylvania State University that contributed to epoch 3, in addition to B6D2 AI progeny (G8 and G9) to make BXD160 through BXD186 contributed by Dr. Abraham Palmer. We thank Dr. Andrew Clark for his helpful discussion.

The UTHSC Center for Integrative and Translational Genomics (CITG) has supported production of the BXD colony at UTHSC and will continue to support this colony for the duration of the grant. The CITG also provides generous support for computer hardware and programming associated with GeneNetwork, and our Galaxy and UCSC Genome Browser instances. We thank the support of the UT Center for Integrative and Translational Genomics and funds from the UT-ORNL Governor's Chair, NIDA grants P30DA044223, P50DA037844 and U01DA051234; NIAAA grants U01AA013499, U01AA016662, and U01AA014425. The BXD Resource at the Jackson Laboratory is supported by NIH P40OD011102 awarded to Dr. Cathleen M. Lutz. The work at Stanford University is supported by NHGRI grant R01HG008140. The data sets generated and/or analyzed during the current study are available in the GeneNetwork repository, https://www.genenetwork.org/.

Methods and materials

Sequencing the BXD families

Tissue was taken from 154 males, mainly young adults (Supplementary Table 1). Mice were euthanized using isoflurane, tissue was collected immediately, flash frozen with liquid nitrogen, and placed in the -80 freezer for later analysis. DNA was extracted from 50 to 80 mg of tissue. DNA extraction, library preparation and sequencing was carried out by HudsonAlpha. High molecular weight (HMW) genomic DNA (gDNA) was isolated using

the Qiagen MagAttract kit. The Chromium Gel Bead and Library Kit (v2 HT kit, revision A; 10X Genomics, Pleasanton, CA, USA) and the Chromium instrument (10X Genomics) were used to prepare the libraries for sequencing. The barcoded libraries were sequenced on an Illumina HiSeq X10 system. The phasing software LongRanger (v2.1.6)[43] was run to generate a phased call-set of single nucleotide variants (SNVs), insertion/deletions (indels), and structural variant discovery, against the mm10 reference genome.

Joint calling used GATK (4.0.3.0). HaplotypeCaller was used to create gvcf files from bam files produced by LongRanger. Variant calls use family-wide information, increasing the likelihood of detecting segregating variants—either between the parental strains, or within the subfamily epochs.

Variant quality was calculated using *variant quality score recalibration* (VQSR) from GATK. A list of known variants was produced by finding those variants we detected in all three of the following independent resources: 1. our own new sequencing of DBA/2J; 2. Wang et al (2016)[18] sequencing of DBA/2J using both SOLiD and Illumina; and 3. the Sanger Mouse Genomes project[28] sequencing of DBA/2J using Illumina. The union of these three includes 3,972,727 SNPs, 404,349 deletions and 365,435 insertions. As expected, these variants segregate with a MAF close to 0.5. This set has been taken as a validated dataset. A training dataset was created using the mgp.v5.merged.snps_all.dbSNP142.vcf.gz (5/12/15) and mgp.v5.merged.indels.dbSNP142.normed.vcf.gz (4/30/15) files from ftp://ftp-mouse.sanger.ac.uk/current_snps.

Structural variant calling

We developed a joint calling method, SVJAM, that detects large structural variants (SV) from linked-read data across multiple samples. A detailed description of the algorithm is available from Gunturkun et al. 2021[42]. Briefly, SVJAM first collects candidate SV regions from individual samples reported by LongRanger[43], which is error prone. We then retrieve barcode-overlapping data for each candidate location from all samples using the Loupe application of the 10x Chromium Platform (Figure 7A, 7B), one image for each individual from a genomic location of interest. The intensity of pixels in these images represent the depth of barcode overlap for the corresponding genomic locations and are the primary data for our analysis.

We used slightly different processes for different types of SV. Conventional image processing techniques are used to detect deletions, which are represented by symmetric gaps with no barcode overlap along the diagonal of the image (Figure 7A, first image). The beginning and end of the deletion are determined according to the location of the gap on the x- and y-axis. Duplication (Figure 7A, second image) and inversion (Figure 7A, third image) are called using a convolutional neural network (CNN) with convolution, max pooling, dropout and flatten layers. This CNN is trained on a set of 12,000 images with validated labels (i.e. type of SV).

To further increase the accuracy of genotype calls, we designed a joint calling procedure that determines the presence of structural variants and the genotype of each individual. This procedure first converts each image into a matrix and then flattens it to a vector. The vectors of each sample are then combined in columns to form a matrix having approximately one million (pixels per image) times 152 (number of samples) dimensions for each candidate location.

The primary algorithm of the joint calling is a principal component (PC) analysis on the high dimensional matrices. Samples projected onto a two-dimensional space of the first two PCs display different patterns based on the number of genotypes: no distinctive pattern is shown when only one genotype is present, two distinctive clusters are shown when there are two genotypes. Because the BXD panel is inbred, when heterozygotic samples are present, they always contain a small number of samples and are located between the two homozygotic clusters (Figure 7C).

We applied hierarchical clustering to the data projected onto the 2D plane to call the genotypes. We designed a custom metric d_M that does not require the samples to be evenly distributed on the two PCs.

$$d_{\mathcal{M}}(p, q) = n \cdot |p_1 - q_1| + |p_2 - q_2|$$

This distance matrix is formed by d_M for each genomic location.

We also built performance evaluation scores for clustering quality and the number of genotypes present. The weight *n* in the above formula is chosen such that it generates the highest clustering quality. Furthermore, we calculate a membership probability that shows the probability of an individual belonging to each genotype cluster. The pipeline then produces a *gvcf* file that captures the information discussed above.

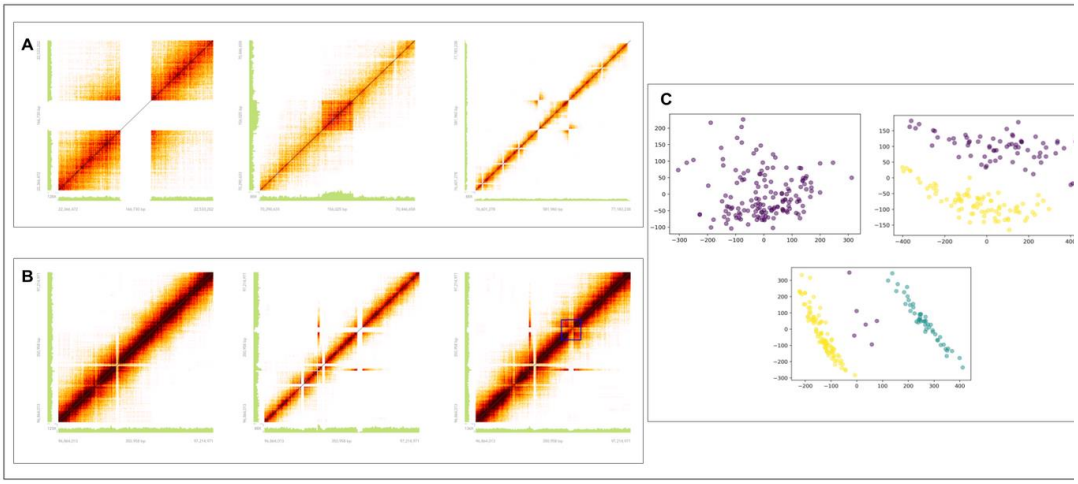


Figure 7: Joint calling of large SVs from linked-read data. (A) Example barcode overlapping images of deletion, duplication and inversion. **(B)** Examples of homozygous reference, homozygous alternative, and heterozygous variants. **(C)** Joint calling showing one, two or three genotypes. Samples (showing as dots) are projected onto a two dimensional space after principal component analysis. Colors represent genotype calls obtained using a clustering algorithm with a custom distance matrix.

Identification of a novel QTL regulating short interspersed element (SINE) transposition rate

To demonstrate our ability to find short interspersed elements (SINEs), we assembled the genome of each strain using SuperNova, blasted the B2 SINE sequence against these 148 assemblies, and counted the number of high quality hits (e-score <1E-10, length >

149 nucleotides, identify > 75%). We then mapped B2 SINE counts using our marker map generated above.

Detection of BXD-private and shared variants

Methods for identifying high-confidence autosomal singleton variants are described in detail in a previous manuscript[35]. Briefly, we used *cyvcf2*[113] to identify single-nucleotide variants at which both founder genotypes (DBA/2J and C57BL/6J) and all but one of the BXD RILs were homozygous for the reference allele; at these sites, we therefore required a single BXD RIL (the focal line) to be either heterozygous or homozygous for the alternate allele. If the focal RIL was heterozygous for the alternate allele, we further required the fraction of reads supporting the alternate allele in that RIL to be >= 0.9. Finally, we required that the genotypes in both of the founders, as well as the focal RIL, were supported by >= 10 reads and had a Phred-scaled genotype quality of at least 20. We also removed all putative singletonsthat overlapped segmental duplications or simple repeat annotations in mm10/GRCm38, which were downloaded from the UCSC Genome Browser[114]. To identify epoch-private variants, we applied the same filters as previously described, but instead required the variant to be present in at least two of the sequenced BXDs, and for all of the BXDs with the shared variant to have the same parental haplotype. We also considered heterozygous genotypes with very high allele balance (i.e., the fraction of reads supporting the alternate allele >= 0.9) to be effectively homozygous. For candidate heterozygous and homozygous singletons (or candidate shared variants), we required the genotype call to be supported by at least 10 sequencing reads and have a Phred-scaled genotype quality of at least 20. Finally, we confirmed that at least one other BXD shared a parental haplotype identical-by-descent with the focal strain (i.e., the strain with the putative singleton) at the singleton site but was homozygous for the reference allele at that site.

We additionally annotated the full autosomal BXD VCF with SnpEff[115] version 4.3t, using the GRCm38.86 database and the following command:

```
java -Xmx16g -jar /path/to/snpeff/jarfile GRCm38.86 /path/to/bxd/vcf > /path/to/uncompressed/output/vcf
```

Structural Variation on the BXD Sex Chromosomes

We estimated copy number states from read depth computed in non-overlapping 500bp sliding windows using *mosdepth*. Specifically, the executed command was:

```
mosdepth -n -b 500 -t 2 -x $PREFIX $BAM
```

Raw read depth values were then corrected for potential GC-biases introduced during library preparation. Briefly, we used the mm10 reference genome to compute the observed GC content of each 500bp window. GC content values were rounded to the nearest 0.001 and regions with identical GC content were binned. For each strain, we then computed the mean read depth across all genomic windows that fell into each GC content bin. Next, we fitted a second degree polynomial to the relationship between read depth and GC content using the *scatter.smooth* function in R and with span parameter = 0.7. For each GC-bin, we then computed the difference between the fitted polynomial and the genome-wide average read depth. These values correspond to the magnitude of "inflation" or "deflation" in read depth across windows of a given GC-content due to systematic GC biases in the data. The read depth value in each 500bp window was then adjusted by the appropriate GC correction factor. Finally, these GC-corrected read depths were divided by the average per-sample coverage to convert into absolute copy number estimates. CN values across Chr X and Chr Y were then visualized using bedGraphs uploaded to IGV.

The length of Chr Y in each genome was inferred from the relative ratio of Chr Y to autosomal mapped reads. Briefly, we obtained the number of sequenced bases on each chromosome from the **.summary.txt* file output by *mosdepth*. We then computed the ratio of the number of sequenced Chr Y bases to the number of autosomal sequenced bases, excluding all unplaced contigs. This ratio was then standardized by the fixed chrY:Autosome ratio for B6 and then multiplied by the size of the Chr Y reference (91,744,698) to convert to bases.

Genome-wide STR genotyping

We used Tandem Repeats Finder[116] (TRF) to identify regions within the mm10 mouse reference genome predicted to harbor STRs with repeat unit lengths up to 20bp using options matchscore=2; mismatchscore=5; indelscore=17; maxperiod=20; pm=80; pi=10; minscore=24; maxlen=1000. We processed this initial reference set using a custom script

to (1) exclude homopolymer repeats which are highly error prone; (2) keep the shortest repeat unit for STRs that share either the same start or end position; (3) exclude "compound" repeats, consisting of multiple directly adjacent STRs with different repeat units; (4) trim repeat regions to only contain perfect repeats with no sequence imperfections; (5) collapse any duplicate STRs introduced by trimming; (6) remove overlapping STRs for which the repeat unit is identical; and (7) exclude very short repeats which we have observed are unlikely to be polymorphic. We filtered dinucleotide STRs with less than 5 perfect copies, trinucleotides with less than 4 perfect copies, and all other repeat classes with less than 3 perfect copies in mm10. After filtering, 1,176,016 STRs remained in our reference.

We used the "fix_read_length" development branch of GangSTR[39] (available at https://github.com/gymreklab/GangSTR/tree/fix_read_length) to genotype the reference STR loci in 152 BXD RI strains and the two founder strains C57BL/6J and DBA/2J from 10X Genomics Illumina short-read sequencing data. To reduce the effects of PCR "stutter" errors, we removed PCR duplicate reads using the –drop-dupes option. We then used HipSTR[117] to estimate per-locus stutter probabilities. We used a custom build to extend HipSTR to ignore the "AS/XS" BAM tags present in Chromium data, which are not properly currently handled in the HipSTR release, and to perform only stutter estimation but not genotyping. We ran HipSTR jointly on BAMs for all strains using our custom reference using option --min-reads 20 to output custom stutter models for each STR. For STRs at which HipSTR could not infer stutter models, including all repeats with repeat units >9bp, we set missing stutter parameters to default values of p=0.9; up=0.05; down=0.05. Additionally, since read length is a critical parameter used in GangSTR's statistical model, we trimmed the second read in each read pair to 128bp to match the length of the first read. We then called GangSTR separately on each strain using our STR reference panel, trimmed and de-dupped reads, and per-locus stutter error probabilities as input. We additionally applied non-default parameter --max-proc-read, which was set to 4500 for DBA/2J which had higher coverage and 3000 for all other strains. This parameter skips loci with extremely high coverage which are likely to be error prone and consume high amounts of memory.

STR genotypes for each strain were filtered using the dumpSTR function from the TRTools package[118] with options --min-call-DP 20; --max-call-DP 1000; --min-call-Q 0.9; --filter-badCI; --require-support 2; --readlen 128 to remove genotype calls with insufficient

read depth, read support, or quality scores. Calls were then merged into a single multi-sample VCF file containing maximum likelihood diploid genotypes for each STR in each strain.

We applied the following filters to remove low-quality STRs from the merged VCF: (1) STRs overlapping known segmental duplication regions in the mm10 reference based on the mm10.genomicSuperDups table obtained from the UCSC Table Browser[114]; (2) STRs with call rates less than 90% across unfiltered strains; (3) STRs with no variation in repeat number across all strains; and (4) STRs for which variants from the mm10 reference were only observed in heterozygous genotypes. The final call-set contained 76,727 STR loci across 152 RI strains with an average per-strain call rate of 96.4%.

Validating STR genotypes using capillary electrophoresis

For each candidate TR, we designed primers to amplify the TR and surrounding region. A universal M13(-21) sequence (5'-TGTAAAACGACGGCCAGT-3') was appended to each forward primer. We then amplified each TR using a three-primer reaction previously described[57] consisting of the forward primer with the M13(-21) sequence, the reverse primer, and a third primer consisting of the M13(-21) sequence labeled with a fluorophore.

The forward (with M13(-21) sequence) and reverse primers for each TR were purchased through IDT. The labeled M13 primers were obtained through ThermoFisher (#450007) with fluorescent labels added to the 5' ends (either FAM, VIC, NED, or PET). TRs were amplified using the forward and reverse primers plus an M13 primer with one of the four fluorophores with GoTaq polymerase (Promega #PRM7123) using PCR program: 94°C for 5 minutes, followed by 30 cycles of 94°C for 30 seconds, 58°C for 45 seconds, 72°C for 45 seconds, followed by 8 cycles of 94°C for 30 seconds, 53°C for 45 seconds, 72°C for 45 seconds, followed by 72°C for 30 minutes.

Fragment analysis of PCR products was performed on a ThermoFisher SeqStudio instrument using the GSLIZ1200 ladder, G5 (DS-33) dye set, and long fragment analysis options. Resulting .fsa files were analyzed using manual review in genemapper.

Pangenome generation, and variant calling

Supernova haployd assemblies obtained by 10X linked reads were mapped against the GRCm38/mm10.fa reference genome using wfmash*v.0.6.0 (https://github.com/ekg/wfmash)* to select for reads mapping to chromosome 19. Mapped assemblies were used to build the pangenome with pggb *(v.0.2.0-pre+d8a5709-2; https://github.com/pangenome/pggb)* using the following combination of parameters:

```
pggb-0.2.0-pre+d8a5709-2 -i chr19.pan+ref.fa.gz -o chr19.pan+ref -t 48 -
p 98 -s 100000 -n 140 -k 229 -O 0.03 -T 20 -U -v -L -Z
```

The variant calling on the pangenome was done using the following combination of parameters in vg[119] *(v1.35.0-59-ge5be425c6).*

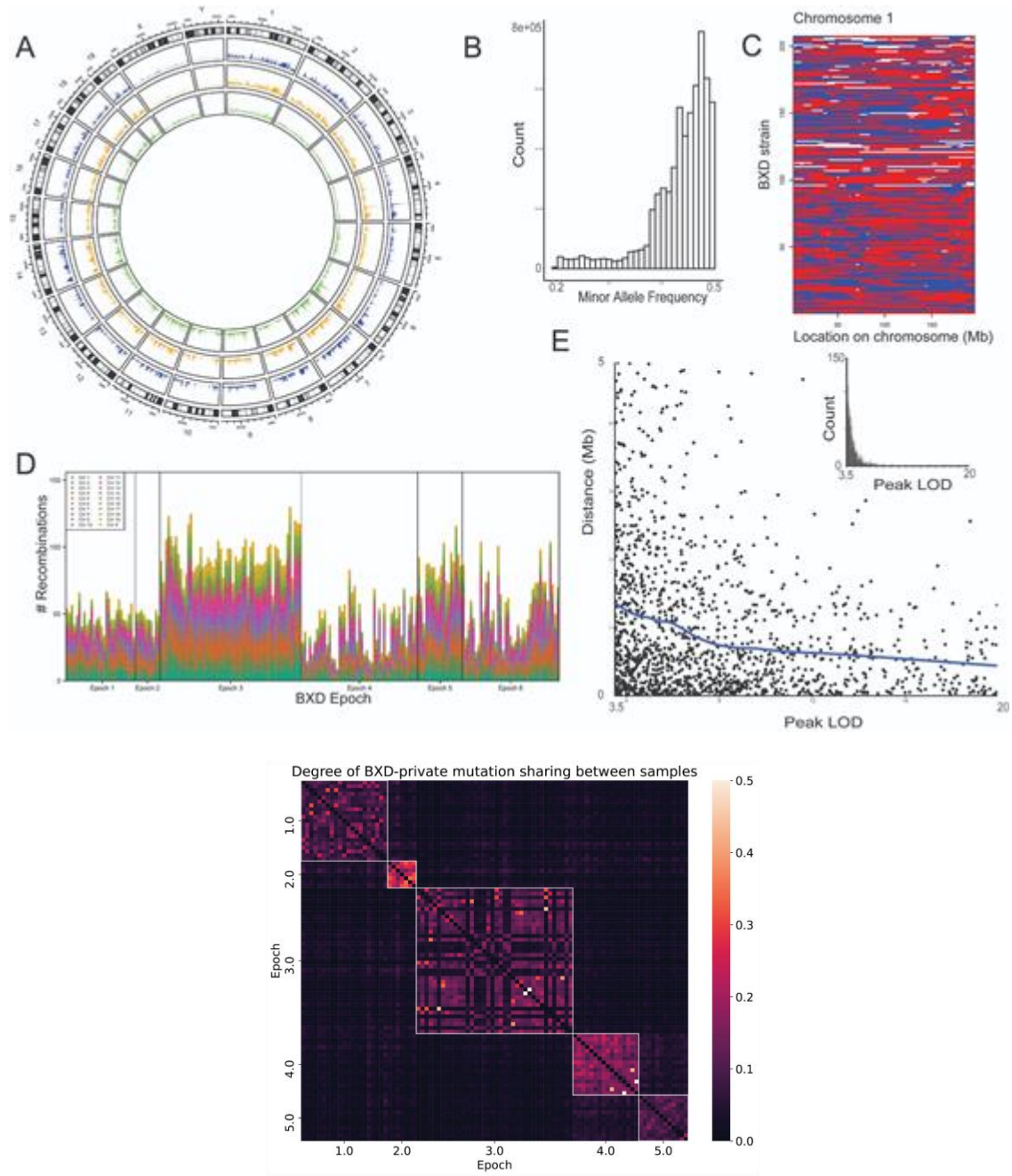
```
vg-e5be425 deconstruct -t 16 -P REF -e -a -H '#' graph.gfa > graph.vcf
```

Validation of the call set was performed on DBA/2J using two true positive sets obtained from GATK *(v.4.0.3.0)* HaplotypeCaller on two sets of raw sequencing: the 10X and the PacBio sequence reads of DBA/2J. Prior to comparison the pangenome-derived, the PacBio-based, and the validated call sets were processed to remove missing data, sites where alleles are stretches of Ns, homozygous reference genotypes and variants greater than 50bp before normalization and decomposition using bcftools[120] under standard parameters. While the pangenome-derived VCF was based on haploid assemblies, for comparison purposes the calls were considered as homozygous diploid in the assumption that DBA/2J is fully isogenic, given ~200 generations of sib-sib inbreeding. Comparison of the three call sets was performed with RTG tools[121] *(v.3.12.1)* using the *--squash-ploidy* option. RepeatMasker[122] was used to mask complex regions.

For pangenome graph visualization we used odgi[123] *(v.0.6.2)* and bandage[124] *(v. 0.8.1)*.

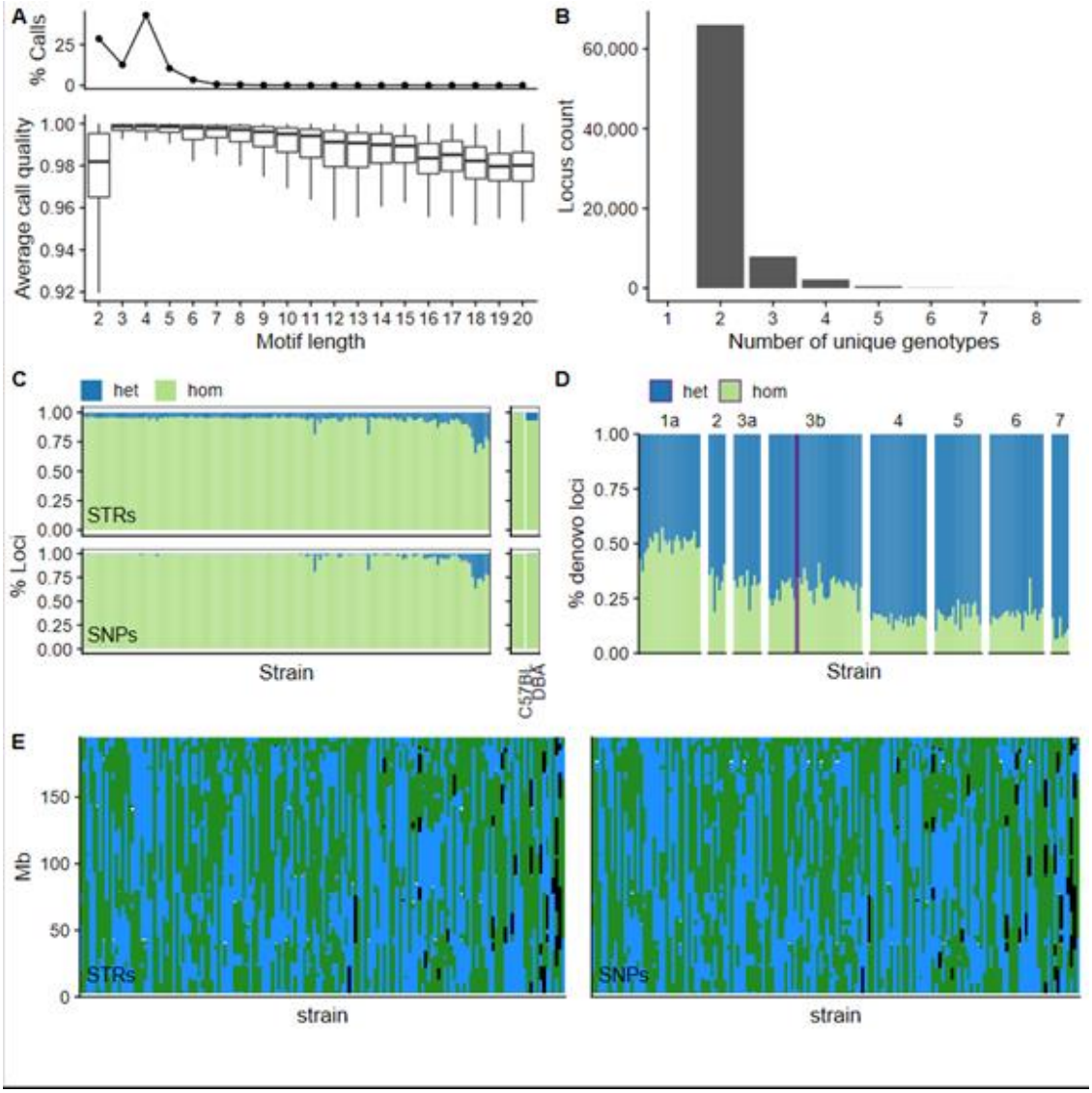
Supplementary information

Supplementary figures

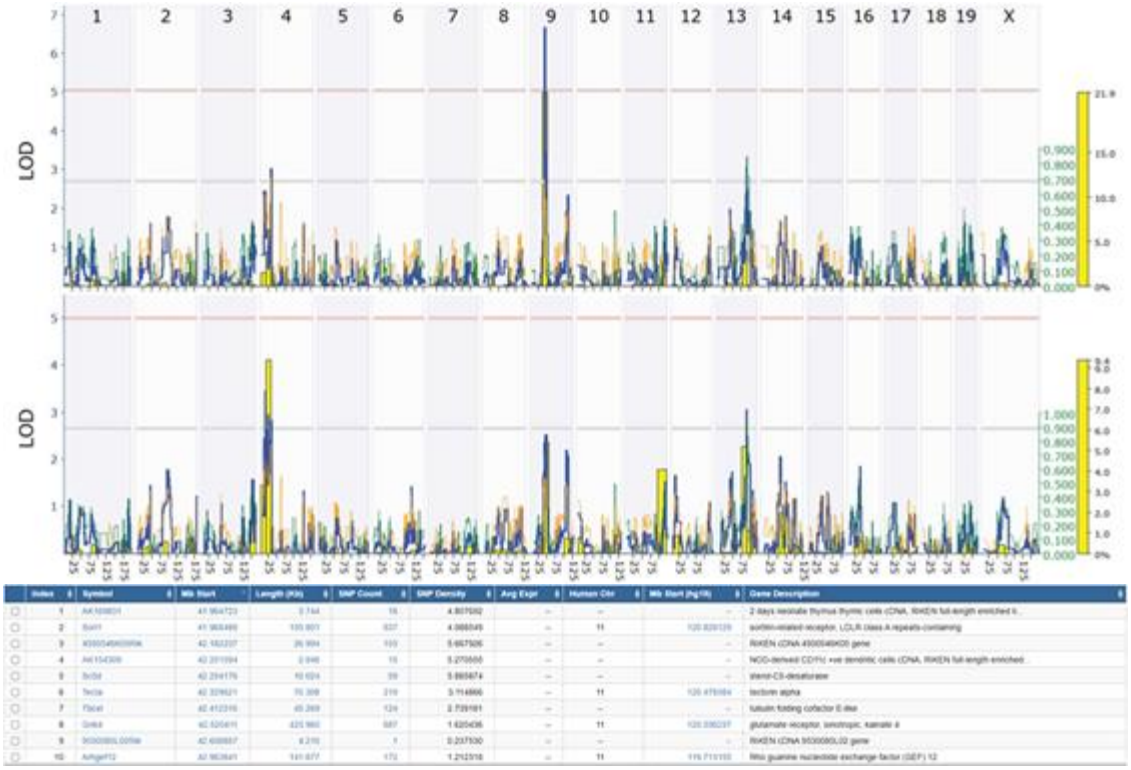


Supplementary Figure 1: Distribution of variants across the genome (A), minor allele frequency (B), chromosome 1 recombination map (C), recombinations per strain (D), and precision of QTL mapping (E). (A) The distribution of SNPs (purple), Indels (orange), and STRs (green) across the BXD genome. Note areas that have a low number of all types of variant, indicating regions that are identical by decent between the C57BL/6J and DBA/2J

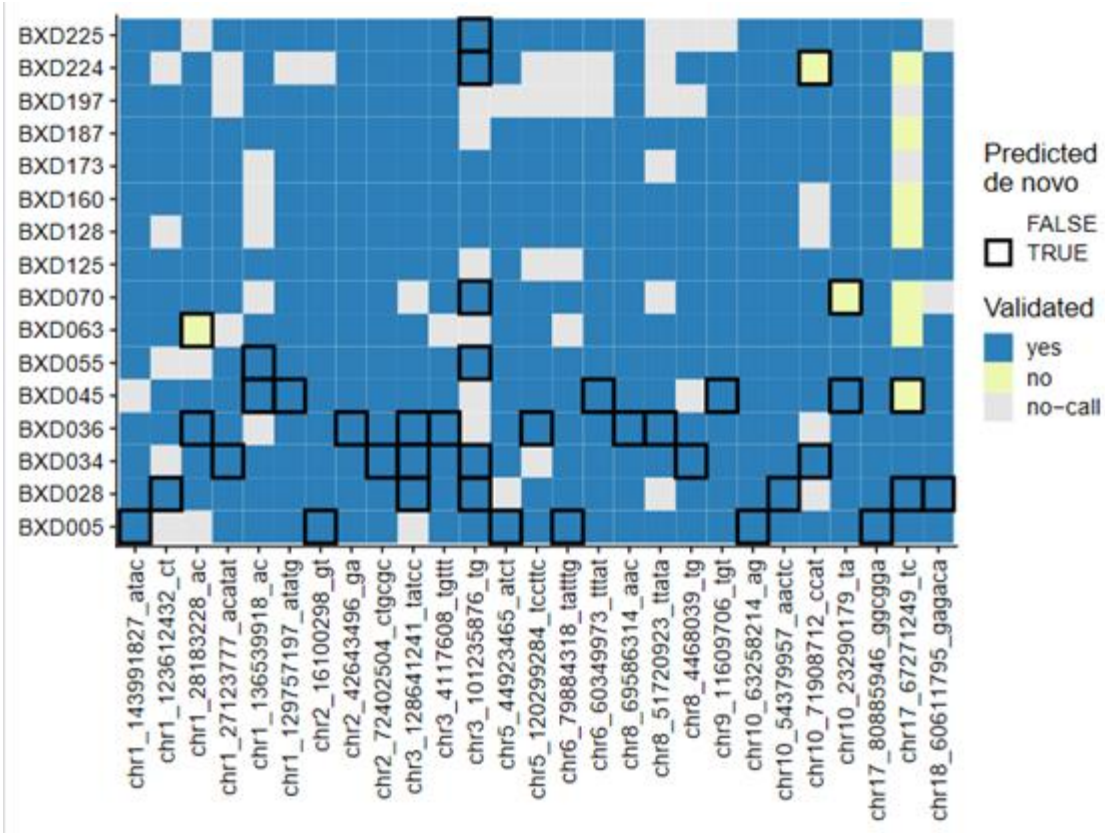
strains. **(B)** The distribution of SNPs with a minor allele frequency (MAF) > 0.2 across the BXD family. The mean MAF is 0.44, which is expected, since the majority of segregating variants will be inherited from the C57BL/6J or DBA/2J parents, and therefore split equally across the population. **(C)** Recombination map of Chr 1. Regions of Chr 1 are shown as either C57BL/6J-like (red), DBA/2J-like (blue) or unknown (white, either due to only having array-based genotypes available or because they were heterozygous at the time of sequencing). **(D)** The number of recombinations on each chromosome per strain, coloured by the chromosome, and divided by the epoch. It is clear to see that epochs 3 and 5, which were produced from advanced intercross lines, have more recombinations than those epochs produced from simple F2s. **(E)** Each point represents the distance from the transcription start site of a gene (TSS) and the peak of a cis-eQTL for that gene, for LOD values between 3.5 (nominally significant) and 20 (highly significant). The blue line is a best fit, showing the mean distance between the TSS and peak LOD. The inset shows a histogram of the number of cis-eQTLs with LODs between 3.5 and 20. **(F)** Mutation sharing between BXD family members. We first identified mutations that were shared by 2 or more members of the BXD family, which likely represent recent *de novo* mutations private to a particular BXD lineage. For every pair of BXD family members, we then counted the number of mutations shared between the pair, and divided that number by the total count of mutations the two BXDs shared with any other BXDs; this fraction represents the degree of pairwise mutation sharing between the two BXD. Each epoch is outlined with a white box to highlight its self-similarity.



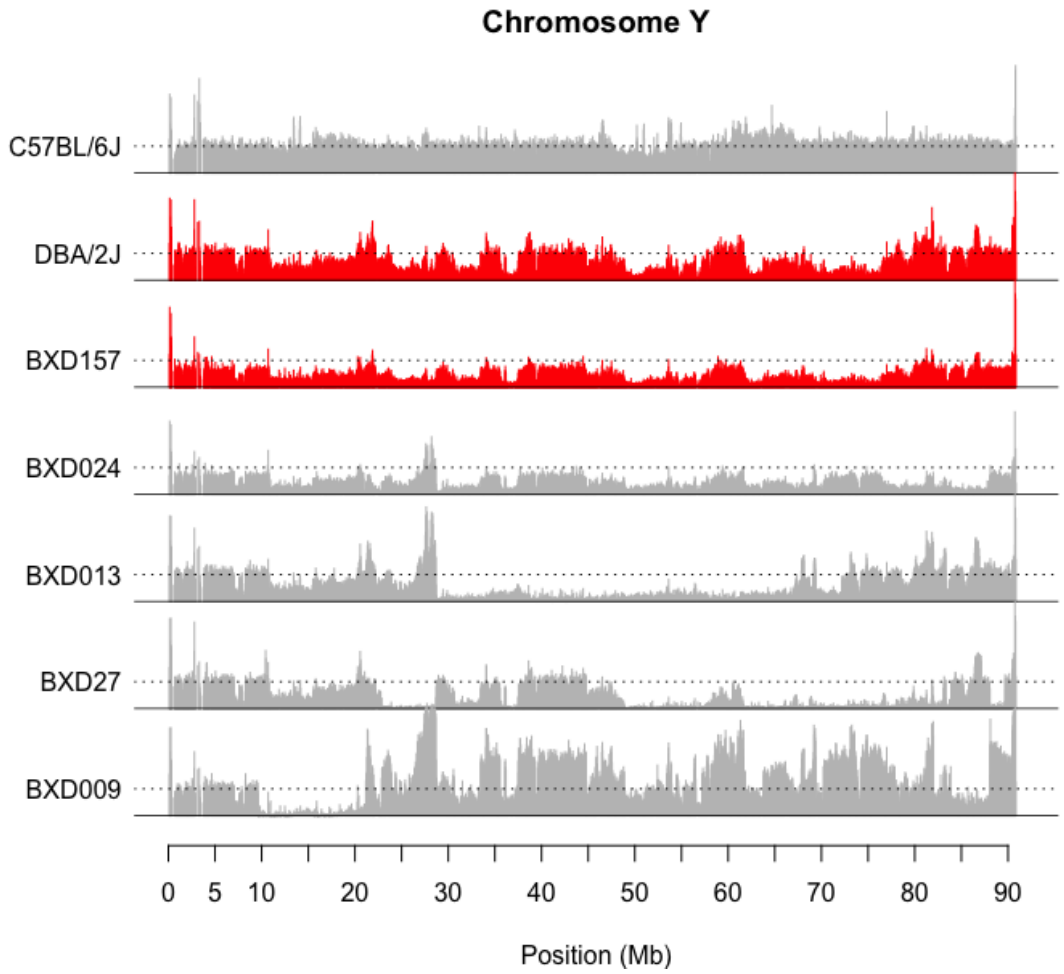
Supplementary Figure 2: Summary of short tandem repeat (STR) identification (A) proportion of (top panel) and average call quality of STRs of a given motif length in the dataset. **(B)** Distribution of the number of unique genotypes per locus for STRs. **(C)** Comparison of the proportion of homozygous (green) and heterozygous (blue) STR calls for each RI strain (left panel) and founders (right panel) between SNPs and STRs. **(D)** Proportion of homozygous to heterozygous *de novo* STR calls for each RI strain. **(E)** Homozygous patchwork of founder haplotype inheritance (C57BL/6J: green; DBA/2J: blue) for BXD RI strains, visualized as a strain by position matrix for STRs (left) and SNPs (right).



Supplementary Figure 3: QTL mapping of water intake of 13-week old females (BXD_12889) using our new genotypes (A) and the old genotypes (B). The higher red horizontal line shows genome-wide significance, whereas the lower grey horizontal line shows the suggestive threshold. Note that the QTL on Chr 9 is not even suggestive in B, but is highly significant in A. The 10 genes within the QTL interval are shown in C.



Supplementary Figure 4. STR genotype validation using capillary fragment analysis. Rows and columns represent strains and loci respectively, which were selected for validation. Heatmap fill indicates validation status: validated (blue), not validated (yellow). Strain/loci combinations missing a GangSTR call to compare to are filled in gray. Strain/loci combinations where a *de novo* STR variant was predicted by GangSTR are highlighted with a black border.



Supplementary Figure 5. Estimated copy number in 500 bp windows across Chr Y in C57BL/6J, DBA/2J, and several BXD strains with extreme Chr Y lengths. The faint dashed horizontal lines correspond to the expected CN=1 copy number state. Strains with DBA/2J-derived Y chromosome are shown in red and those with a C57BL/6J-derived Chr Y are depicted in gray.

.

Supplementary Tables

Supplementary Table 1
https://docs.google.com/spreadsheets/d/16WzQc1qM-ehDar8UPmVVQArr41QTI5i54aMVVsDm8Kg/edit?usp=sharing

Supplementary Table 2: Variable sites in DBA/2J as called by GATK on genomics versus vg on pangenomic data. Truth sets are GATK on 10X (GATK-10X) and GATK on PacBio (GATK-PacBio). Only microvariants (length up to 50bp) are considered.

Type of microvariant	Type of variant	Number of variable sites		
		GATK-10X	GATK-PacBio	vg-10X
simple	**SNPs**	143,060	155,426	132,963
simple	**INDELS**	60,700	52,264	31,771
simple	**MNPs**	0	0	13,823
	total simple	***203,760***	***207,690***	***178,557***
complex	**SNP/INDEL**	2,576	269	1,255
complex	**MNP/CLUMPED**	0	0	260
complex	**MNP/INDEL**	0	0	1925

complex	**INDEL/CLUMPED**	0	0	646
	total complex	***2,576***	***269***	***4,086***
Total variable sites		**206,336**	**207,959**	**182,643**

Supplementary Table 3: Quality of variant calling from the pangenome in DBA/2J. GK-10x: GATK on 10X sequence data; GK-PB: GATK on PacBio sequence data; vg-10X: vg on 10X sequence data. Gray columns (GK-10X vs GK-PB) are comparisons among the truth sets, in this case only the sequence technology is being evaluated.

		PRECISION (%)			**SENSITIVITY (%)**			**F-MEASURE (%)**		
		GK-10X *vs* GK-PB	**GK-10X *vs* vg-10X**	**GK-PB *vs* vg-10X**	**GK-10X *vs* GK-PB**	**GK-10X *vs* vg-10X**	**GK-PB *vs* vg-10X**	**GK-10X *vs* GK-PB**	**GK-10X *vs* vg-10X**	**GK-PB *vs* vg-10X**
Masked	**All**	92	90	90	91	85	86	92	87	88
	SNPs	95	93	94	97	85	84	96	89	89
	Indels	79	88	86	70	52	57	74	65	69
Unmasked	**All**	77	75	79	82	71	70	79	73	74

Bibliography

1. Cubillos, F. A. *et al.* Assessing the complex architecture of polygenic traits in diverged yeast populations. *Mol. Ecol.* **20**, 1401–1413 (2011).
2. El-Din El-Assal, S., Alonso-Blanco, C., Peeters, A. J. M., Raz, V. & Koornneef, M. A QTL for flowering time in Arabidopsis reveals a novel allele of CRY2. *Nat. Genet.* **29**, 435–440 (2001).
3. Lister, C. & Dean, C. Recombinant inbred lines for mapping RFLP and phenotypic markers in Arabidopsis thaliana. *Plant J.* **4**, 745–750 (1993).
4. Pan, Q. *et al.* The genetic basis of plant architecture in 10 maize recombinant inbred line populations. *Plant Physiol.* **175**, 858–873 (2017).
5. Yin, X., Struik, P. C., Tang, J., Qi, C. & Liu, T. Model analysis of flowering

phenology in recombinant inbred lines of barley. *J. Exp. Bot.* **56**, 959–965 (2005).
6. Ruden, D. M. *et al.* Genetical toxicogenomics in Drosophila identifies master-modulatory loci that are regulated by developmental exposure to lead. *Neurotoxicology* **30**, 898–914 (2009).
7. Cochrane, B. J., Windelspecht, M., Brandon, S., Morrow, M. & Dryden, L. Use of recombinant inbred lines for the investigation of insecticide resistance and cross resistance in Drosophila simulans. *Pestic. Biochem. Physiol.* **61**, 95–114 (1998).
8. Snoek, B. L. *et al.* A multi-parent recombinant inbred line population of C. elegans allows identification of novel QTLs for complex life history traits. *BMC Biol.* **17**, 24 (2019).
9. Fitzgerald, T. *et al.* The Medaka Inbred Kiyosu-Karlsruhe (MIKK) Panel. *bioRxiv* 2021.05.17.444412 (2021) doi:10.1101/2021.05.17.444412.
10. Printz, M. P., Jirout, M., Jaworski, R., Alemayehu, A. & Kren, V. Genetic models in applied physiology. HXB/BXH rat recombinant inbred strain platform: a newly enhanced tool for cardiovascular, behavioral, and developmental genetics and genomics. *J. Appl. Physiol.* **94**, 2510–2522 (2003).
11. Ashbrook, D. G. *et al.* A platform for experimental precision medicine: The extended BXD mouse family. *Cell Syst.* **12**, 235-247.e9 (2021).
12. Crow, J. F. Haldane, Bailey, Taylor and recombinant-inbred lines. *Genetics* **176**, 729–732 (2007).
13. Bailey, D. W. Recombinant-inbred strains. An aid to finding identity, linkage, and function of histocompatibility and other genes. *Transplantation* **11**, 325–327 (1971).
14. Taylor, B. A., Heiniger, H. J. & Meier, H. Genetic analysis of resistance to cadmium-induced testicular damage in mice. *Proc. Soc. Exp. Biol. Med.* **143**, 629–633 (1973).
15. Taylor, B. A. *et al.* Genotyping new BXD recombinant inbred mouse strains and comparison of BXD and consensus maps. *Mamm. Genome* **10**, 335–348 (1999).
16. Sandoval-Sierra, J. V. *et al.* Body weight and high-fat diet are associated with epigenetic aging in female members of the BXD murine family. *Aging Cell* e13207 (2020) doi:10.1111/acel.13207.
17. Peirce, J. L., Lu, L., Gu, J., Silver, L. M. & Williams, R. W. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC Genet.* **5**, 7 (2004).
18. Wang, X. *et al.* Joint mouse-human phenome-wide association to test gene function and disease risk. *Nat. Commun.* **7**, 10464 (2016).
19. Roy, S. *et al.* Gene-by-environment modulation of lifespan and weight gain in the murine BXD family. *Nat. Metab.* **3**, 1217–1227 (2021).
20. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. data* **3**, 160018 (2016).
21. Ashbrook, D. G., Mulligan, M. K. & Williams, R. W. Post-genomic behavioral genetics: From revolution to routine. *Genes. Brain. Behav.* **17**, e12441 (2018).
22. Sloan, Z. *et al.* GeneNetwork: framework for web-based genetics. *J. Open Source Softw.* **1**, 25 (2016).
23. Mulligan, M. K., Mozhui, K., Prins, P. & Williams, R. W. GeneNetwork: A Toolbox for Systems Genetics. *Methods Mol. Biol.* **1488**, 75–120 (2017).
24. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
25. Church, D. M. *et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* **7**, e1000112 (2009).
26. Lloyd, K., Franklin, C., Lutz, C. & Magnuson, T. Reproducibility: Use mouse

biobanks or lose them. *Nature* **522**, 151–153 (2015).
27. Doran, A. G. *et al.* Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations. *Genome Biol.* **17**, 167 (2016).
28. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
29. Wang, X. *et al.* High-throughput sequencing of the DBA/2J mouse genome. *BMC Bioinformatics* **11**, O7 (2010).
30. Li, H. & Auwerx, J. Mouse systems genetics as a prelude to precision medicine. *Trends Genet.* **36**, 259–272 (2020).
31. Yalcin, B., Adams, D. J., Flint, J. & Keane, T. M. Next-generation sequencing of experimental mouse strains. *Mamm. Genome* **23**, 490–498 (2012).
32. Lindsay, S. J., Rahbari, R., Kaplanis, J., Keane, T. & Hurles, M. E. Similarities and differences in patterns of germline mutation between mice and humans. *Nat. Commun.* **10**, 4053 (2019).
33. Srivastava, A. *et al.* Genomes of the mouse collaborative cross. *Genetics* **206**, 537–556 (2017).
34. Shorter, J. R. *et al.* Whole genome wequencing and progress toward full inbreeding of the mouse Collaborative Cross population. *G3 (Bethesda).* **9**, 1303–1311 (2019).
35. Sasani, T. A. *et al.* A wild-derived antimutator drives germline mutation spectrum differences in a genetically diverse murine family. *bioRxiv* 2021.03.12.435196 (2021) doi:10.1101/2021.03.12.435196.
36. Maksimov, M. *et al.* A novel quantitative trait locus implicates Msh3 in the propensity for genome-wide short tandem repeat expansions in mice. *bioRxiv* 2022.03.02.482700 (2022) doi:10.1101/2022.03.02.482700.
37. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1-11.10.33 (2013).
38. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
39. Mousavi, N., Shleizer-Burko, S., Yanicky, R. & Gymrek, M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 (2019).
40. Willems, T. *et al.* Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* **14**, 590–592 (2017).
41. Clarke, L. A., Rebelo, C. S., Gonçalves, J., Boavida, M. G. & Jordan, P. PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. *Mol. Pathol.* **54**, 351–353 (2001).
42. Gunturkun, M. H. *et al.* SVJAM: Joint analysis of structural variants using linked read sequencing data. *bioRxiv* 2021.11.02.467006 (2021) doi:10.1101/2021.11.02.467006.
43. Marks, P. *et al.* Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res.* **29**, 635–645 (2019).
44. Taft, R. A., Davisson, M. & Wiles, M. V. Know thy mouse. *Trends Genet.* **22**, 649–653 (2006).
45. Sarsani, V. K. *et al.* The genome of C57BL/6J 'Eve', the mother of the laboratory mouse genome reference strain. *G3 (Bethesda).* **9**, 1795–1805 (2019).
46. Mortazavi, M. *et al.* Polymorphic SNPs, short tandem repeats and structural variants are responsible for differential gene expression across C57BL/6 and C57BL/10 substrains. *bioRxiv* 2020.03.16.993683 (2021)

doi:10.1101/2020.03.16.993683.
47. Watanabe, T., Cheng, E., Zhong, M. & Lin, H. Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline. *Genome Res.* **25**, 368–380 (2015).
48. Chang, B. *et al.* In-frame deletion in a novel centrosomal/ciliary protein CEP290/NPHP6 perturbs its interaction with RPGR and results in early-onset retinal degeneration in the rd16 mouse. *Hum. Mol. Genet.* **15**, 1847–1857 (2006).
49. Seecharan, D. J., Kulkarni, A. L., Lu, L., Rosen, G. D. & Williams, R. W. Genetic control of interconnected neuronal populations in the mouse primary visual system. *J. Neurosci.* **23**, 11178–11188 (2003).
50. Datta, P., Hendrickson, B., Brendalen, S., Ruffcorn, A. & Seo, S. The myosin-tail homology domain of centrosomal protein 290 is essential for protein confinement between the inner and outer segments in photoreceptors. *J. Biol. Chem.* **294**, 19119–19136 (2019).
51. Rosen, G. D. *et al.* Bilateral subcortical heterotopia with partial callosal agenesis in a mouse mutant. *Cereb. Cortex* **23**, 859–872 (2013).
52. Truong, D. T., Bonet, A., Rendall, A. R., Rosen, G. D. & Fitch, R. H. A behavioral evaluation of sex differences in a mouse model of severe neuronal migration disorder. *PLoS One* **8**, e73144 (2013).
53. Bi, W. *et al.* Increased LIS1 expression affects human and mouse brain development. *Nat. Genet.* **41**, 168–177 (2009).
54. Katayama, K.-I., Hayashi, K., Inoue, S., Sakaguchi, K. & Nakajima, K. Enhanced expression of Pafah1b1 causes over-migration of cerebral cortical neurons into the marginal zone. *Brain Struct. Funct.* **222**, 4283–4291 (2017).
55. Haverfield, E. V, Whited, A. J., Petras, K. S., Dobyns, W. B. & Das, S. Intragenic deletions and duplications of the LIS1 and DCX genes: a major disease-causing mechanism in lissencephaly and subcortical band heterotopia. *Eur. J. Hum. Genet.* **17**, 911–918 (2009).
56. Dobyns, W. B. & Das, S. *PAFAH1B1-Associated Lissencephaly/Subcortical Band Heterotopia*. *GeneReviews®* (1993).
57. Bryant, C. D. *et al.* Facilitating complex trait analysis via reduced complexity crosses. *Trends Genet.* **36**, 549–562 (2020).
58. Eizenga, J. M. *et al.* Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).
59. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.* **19**, 118–135 (2018).
60. Sirén, J. *et al.* Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).
61. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
62. Wang, F. *et al.* ZFP91 disturbs metabolic fitness and antitumor activity of tumor-infiltrating T cells. *J. Clin. Invest.* **131**, (2021).
63. Wang, A. *et al.* ZFP91 is required for the maintenance of regulatory T cell homeostasis and function. *J. Exp. Med.* **218**, (2021).
64. Théberge, E. T. *et al.* Genetic influences on the amount of cell death in the neural tube of BXD mice exposed to acute ethanol at midgestation. *Alcohol. Clin. Exp. Res.* **43**, 439–452 (2019).
65. Zhou, D. *et al.* Ethanol's effect on Coq7 expression in the hippocampus of mice. *Front. Genet.* **9**, 602 (2018).
66. Mulligan, M. K. *et al.* Genetic contribution to initial and progressive alcohol intake among recombinant inbred strains of mice. *Front. Genet.* **9**, 370 (2018).

67. Dickson, P. E. *et al.* Systems genetics of sensation seeking. *Genes. Brain. Behav.* **18**, e12519 (2019).
68. Wang, L. S. *et al.* Critical evaluation of transcription factor Atf2 as a candidate modulator of alcohol preference in mouse and human populations. *Genet. Mol. Res.* **12**, 5992–6005 (2013).
69. Chella Krishnan, K. *et al.* Genetic architecture of group a streptococcal necrotizing soft tissue infections in the mouse. *PLoS Pathog.* **12**, e1005732 (2016).
70. Russo, L. M., Abdeltawab, N. F., O'Brien, A. D., Kotb, M. & Melton-Celsa, A. R. Mapping of genetic loci that modulate differential colonization by Escherichia coli O157:H7 TUV86-2 in advanced recombinant inbred BXD mice. *BMC Genomics* **16**, 947 (2015).
71. Boon, A. C. M., Williams, R. W., Sinasac, D. S. & Webby, R. J. A novel genetic locus linked to pro-inflammatory cytokines after virulent H5N1 virus infection in mice. *BMC Genomics* **15**, 1017 (2014).
72. Nedelko, T. *et al.* Distinct gene loci control the host response to influenza H1N1 virus infection in a time-dependent manner. *BMC Genomics* **13**, 411 (2012).
73. Boon, A. C. M. *et al.* Host genetic variation affects resistance to infection with a highly pathogenic H5N1 influenza A virus in mice. *J. Virol.* **83**, 10417–10426 (2009).
74. Rodrigues, B. de A. *et al.* Obesity increases mitogen-activated protein kinase phosphatase-3 levels in the hypothalamus of mice. *Front. Cell. Neurosci.* **11**, 313 (2017).
75. Jha, P. *et al.* Systems analyses reveal physiological roles and genetic regulators of liver lipid species. *Cell Syst.* **6**, 722-733.e6 (2018).
76. Jha, P. *et al.* Genetic regulation of plasma lipid species and their association with metabolic phenotypes. *Cell Syst.* **6**, 709-721.e6 (2018).
77. Jones, B. C. & Jellen, L. C. Systems genetics analysis of iron and its regulation in brain and periphery. *Methods Mol. Biol.* **1488**, 467–480 (2017).
78. Reyes Fernandez, P. C., Replogle, R. A., Wang, L., Zhang, M. & Fleet, J. C. Novel genetic loci control calcium absorption and femur bone mass as well as their response to low calcium intake in male BXD recombinant inbred mice. *J. Bone Miner. Res.* **31**, 994–1002 (2016).
79. Fleet, J. C. *et al.* Gene-by-Diet interactions affect serum 1,25-Dihydroxyvitamin D levels in male BXD recombinant inbred mice. *Endocrinology* **157**, 470–81 (2016).
80. Diessler, S. *et al.* A systems genetics resource and analysis of sleep regulation in the mouse. *PLoS Biol.* **16**, e2005750 (2018).
81. Jung, S. H., Brownlow, M. L., Pellegrini, M. & Jankord, R. Divergence in Morris Water Maze-based cognitive performance under chronic stress Is associated with the hippocampal whole transcriptomic modification in mice. *Front. Mol. Neurosci.* **10**, 275 (2017).
82. King, R., Lu, L., Williams, R. W. & Geisert, E. E. Transcriptome networks in the mouse retina: An exon level BXD RI database. *Mol. Vis.* **21**, 1235–1251 (2015).
83. Li, H. *et al.* An integrated systems genetics and omics toolkit to probe gene function. *Cell Syst.* **6**, 90-102.e4 (2018).
84. Parsons, M. J. *et al.* Genetic variation in hippocampal microRNA expression differences in C57BL/6 J X DBA/2 J (BXD) recombinant inbred mouse strains. *BMC Genomics* **13**, 476 (2012).
85. Mulligan, M. K. *et al.* Expression, covariation, and genetic regulation of miRNA Biogenesis genes in brain supports their role in addiction, psychiatric disorders, and disease. *Front. Genet.* **4**, 126 (2013).
86. Williams, E. G. *et al.* Quantifying and localizing the mitochondrial proteome across

five tissues in a mouse population. *Mol. Cell. Proteomics* **17**, 1766–1777 (2018).
87. Williams, E. G. *et al.* Multiomic profiling of the liver across diets and age in a diverse mouse population. *Cell Syst.* (2021) doi:10.1016/j.cels.2021.09.005.
88. Williams, E. G. *et al.* Systems proteomics of liver mitochondria function. *Science* **352**, aad0189 (2016).
89. Wu, Y. *et al.* Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. *Cell* **158**, 1415–1430 (2014).
90. Baker, C. L. *et al.* Tissue-specific trans regulation of the mouse epigenome. *Genetics* **211**, 831–845 (2019).
91. Perez-Munoz, M. E. *et al.* Diet modulates cecum bacterial diversity and physiological phenotypes across the BXD mouse genetic reference population. *PLoS One* **14**, e0224100 (2019).
92. McKnite, A. M. *et al.* Murine gut microbiota is defined by host genetics and modulates variation of metabolic traits. *PLoS One* **7**, e39191 (2012).
93. Li, Z. *et al.* A transposon in Comt generates mRNA variants and causes widespread expression and behavioral differences among mice. *PLoS One* **5**, e12181 (2010).
94. Williams, R. *et al.* A common and unstable copy number variant is associated with differences in Glo1 expression and anxiety-like behavior. *PLoS One* **4**, e4649 (2009).
95. Chunduri, A. & Ashbrook, D. G. Old data and friends improve with age: Advancements with the updated tools of GeneNetwork. *bioRxiv* 2021.05.24.445383 (2021) doi:10.1101/2021.05.24.445383.
96. Parker, C. C., Dickson, P. E., Philip, V. M., Thomas, M. & Chesler, E. J. Systems genetic analysis in GeneNetwork.org. *Curr. Protoc. Neurosci.* **79**, 8.39.1-8.39.20 (2017).
97. Watson, P. M. & Ashbrook, D. G. GeneNetwork: a continuously updated tool for systems genetics analyses. *bioRxiv* 2020.12.23.424047 (2020) doi:10.1101/2020.12.23.424047.
98. Ashbrook, D. G. & Lu, L. Recombinant Inbred Mice as Models for Experimental Precision Medicine and Biology. in *Animal Models in Medicine and Biology* (ed. Purevjav, E.) In Press (IntechOpen, 2021). doi:10.5772/intechopen.96173.
99. Collaborative Cross Consortium. The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics* **190**, 389–401 (2012).
100. Threadgill, D. W., Miller, D. R., Churchill, G. A. & de Villena, F. P.-M. The collaborative cross: a recombinant inbred mouse population for the systems genetic era. *ILAR J.* **52**, 24–31 (2011).
101. Threadgill, D. W. & Churchill, G. A. Ten years of the collaborative cross. *Genetics* **190**, 291–294 (2012).
102. Chesler, E. J. *et al.* The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mamm. Genome* **19**, 382–389 (2008).
103. Churchill, G. a *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat. Genet.* **36**, 1133–1137 (2004).
104. Dickinson, M. E. *et al.* High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508–514 (2016).
105. International Mouse Knockout Consortium, Collins, F. S., Rossant, J. & Wurst, W. A mouse for all reasons. *Cell* **128**, 9–13 (2007).
106. Collins, F. S., Finnell, R. H., Rossant, J. & Wurst, W. A new partner for the international knockout mouse consortium. *Cell* **129**, 235 (2007).
107. Neuner, S. M. *et al.* Systems genetics identifies Hp1bp3 as a novel modulator of

cognitive aging. *Neurobiol. Aging* **46**, 58–67 (2016).
108. Neuner, S. M., Heuer, S. E., Huentelman, M. J., O'Connell, K. M. S. & Kaczorowski, C. C. Harnessing genetic complexity to enhance translatability of Alzheimer's disease mouse models: A path toward precision medicine. *Neuron* **101**, 399-411.e5 (2019).
109. O'Connell, K. M. S., Ouellette, A. R., Neuner, S. M., Dunn, A. R. & Kaczorowski, C. C. Genetic background modifies CNS-mediated sensorimotor decline in the AD-BXD mouse model of genetic diversity in Alzheimer's disease. *Genes. Brain. Behav.* **18**, e12603 (2019).
110. Neuner, S. M., Heuer, S. E., Zhang, J.-G., Philip, V. M. & Kaczorowski, C. C. Identification of pre-symptomatic gene signatures that predict resilience to cognitive decline in the genetically diverse AD-BXD model. *Front. Genet.* **10**, 35 (2019).
111. Cowin, R.-M. *et al.* Genetic background modulates behavioral impairments in R6/2 mice and suggests a role for dominant genetic modifiers in Huntington's disease pathogenesis. *Mamm. Genome* **23**, 367–377 (2012).
112. Sipe, L. M. *et al.* Abstract 2919: Novel pre-clinical model to identify genetic modifiers of triple negative breast cancer. in *Tumor Biology* (American Association for Cancer Research, 2021). doi:10.1158/1538-7445.AM2021-2919.
113. Pedersen, B. S. & Quinlan, A. R. cyvcf2: fast, flexible variant analysis with Python. *Bioinformatics* **33**, 1867–1869 (2017).
114. Navarro Gonzalez, J. *et al.* The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.* **49**, D1046–D1057 (2021).
115. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* **6**, 80–92 (2012).
116. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
117. Valle-Silva, G. *et al.* Analysis and comparison of the STR genotypes called with HipSTR, STRait Razor and toaSTR by using next generation sequencing data in a Brazilian population sample. *Forensic Sci. Int. Genet.* **58**, 102676 (2022).
118. Mousavi, N. *et al.* TRTools: a toolkit for genome-wide analysis of tandem repeats. *Bioinformatics* **37**, 731–733 (2021).
119. Hickey, G. *et al.* Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
120. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
121. Cleary, J. G. *et al.* Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *bioRxiv* 23754 (2015) doi:10.1101/023754.
122. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **Chapter 4**, Unit 4.10 (2009).
123. Guarracino, A., Heumos, S., Nahnsen, S., Prins, P. & Garrison, E. ODGI: understanding pangenome graphs. *bioRxiv* 2021.11.10.467921 (2021) doi:10.1101/2021.11.10.467921.
124. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).