Content and Performance of the MiniMUGA Genotyping Array: A New Tool To Improve Rigor and Reproducibility in Mouse Research

John Sebastian Sigmon,*,[1] Matthew W. Blanchard,[†,‡,1] Ralph S. Baric,[§] Timothy A. Bell,[†] Jennifer Brennan[‡]
Gudrun A. Brockmann,** A. Wesley Burks,[††] J. Mauro Calabrese,[‡‡,§§] Kathleen M. Caron,***
Richard E. Cheney,*** Dominic Ciavatta,[†] Frank Conlon,[†††] David B. Darr,[§§] James Faber,*** Craig Franklin,[‡‡‡]
Timothy R. Gershon,[§§§] Lisa Gralinski,[§] Bin Gu,*** Christiann H. Gaines,[†] Robert S. Hagan,****
Ernest G. Heimsath,[§§,***] Mark T. Heise,[†] Pablo Hock,[†] Folami Ideraabdullah,[†,§§,††††] J. Charles Jennette,[‡‡‡‡]
Tal Kafri,[§§§§,*****] Anwica Kashfeen,* Mike Kulis,[††] Vivek Kumar,[†††††] Colton Linnertz,[†] Alessandra Livraghi-
Butrico,[‡‡‡‡‡] K. C. Kent Lloyd,[§§§§§,******,††††††] Cathleen Lutz,[†††††] Rachel M. Lynch,[†,§§] Terry Magnuson,[†,‡,§§]
Glenn K. Matsushima[§§§§,‡‡‡‡‡‡] Rachel McMullan,[†] Darla R. Miller,[†,§§] Karen L. Mohlke,[†] Sheryl S. Moy,[§§§§§§,*******]
Caroline E. Y. Murphy,[†] Maya Najarian,* Lori O'Brien,*** Abraham A. Palmer,[†††††††] Benjamin D. Philpot,***,[‡‡‡‡‡]
Scott H. Randell,*** Laura Reinholdt,[†††††] Yuyu Ren,[††††††] Steve Rockwood,[†††††] Allison R. Rogala,[‡‡‡‡,‡‡‡‡‡‡‡]
Avani Saraswatula,[†] Christopher M. Sassetti,[§§§§§§§] Jonathan C. Schisler,[‡‡] Sarah A. Schoenrock,[†] Ginger D. Shaw[†]
John R. Shorter,[†] Clare M. Smith,[§§§§§§§] Celine L. St. Pierre,[†††††††] Lisa M. Tarantino,[†,********]
David W. Threadgill,[†††††††,‡‡‡‡‡‡‡‡] William Valdar,[†] Barbara J. Vilen,[§§§§] Keegan Wardwell,[†††††]
Jason K. Whitmire,[†] Lucy Williams,[†] Mark J. Zylka,*** Martin T. Ferris,[†,1,2] Leonard McMillan,*,[1]
and Fernando Pardo Manuel de Villena[†,‡,§§,1]

*Department of Computer Science, [†]Department of Genetics, [‡]Mutant Mouse Resource and Research Center, [§]Department of Epidemiology, Gillings School of Public Health, [††]Department of Pediatrics, [‡‡]Department of Pharmacology,[§§]Lineberger Comprehensive Cancer Center, ***Department of Cell Biology and Physiology,[†††]Department of Biology,[§§§]Department of Neurology, ****Division of Pulmonary Diseases and Critical Care Medicine,[††††]Department of Nutrition, Gillings School of Public Health, [‡‡‡‡]Department of Pathology and Laboratory Medicine,[§§§§]Department of Microbiology and Immunology, *****Gene Therapy Center,[‡‡‡‡‡]Marsico Lung Institute/UNC Cystic Fibrosis Center,[§§§§§§]Department of Psychiatry, [‡‡‡‡‡‡]UNC Neuroscience Center, *******Carolina Institute for Developmental Disabilities,[‡‡‡‡‡‡‡]Division of Comparative Medicine, and ********Division of Pharmacotherapy and Experimental Therapeutics, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599, **Humbolt University of Berlin, Berlin, Germany 10117, [‡‡‡]Department of Veterinary Pathobiology, University of Missouri, Columbia, Missouri 65211, [†††††]The Jackson Laboratory, Bar Harbor, Maine 04609, [§§§§§]Department of Surgery, ******School of Medicine, and [††††††]Mouse Biology Program, University of California Davis, California 95616, [†††††††]University of California San Diego, La Jolla, California 92093, [§§§§§§§]Department of Microbiology and Physiological Systems, University of Massachusetts Medical School, Worcester, Massachusetts 01655, and [††††††††]Department of Molecular and Cellular Medicine and [‡‡‡‡‡‡‡‡]Department of Biochemistry and Biophysics, Texas A&M University, Texas 77843

ORCID IDs: 0000-0003-3977-2115 (M.W.B.); 0000-0002-4387-2947 (G.A.B.); 0000-0002-1504-0086 (R.S.H.); 0000-0003-2969-8193 (C.L.); 0000-0001-8164-0744 (R.M.L.); 0000-0002-0792-835X (T.M.); 0000-0002-0781-7254 (D.R.M.); 0000-0001-09886-3178 (C.E.Y.M.); 0000-0003-3634-0747 (A.A.P.); 0000-0003-4054-4048 (L.R.); 0000-0001-7382-2783 (J.C.S.); 0000-0003-4732-5526 (J.R.S.); 0000-0001-5465-6601 (C.L.S.P.); 0000-0002-2419-0430 (W.V.); 0000-0003-1241-6268 (M.T.F.); 0000-0002-5738-5795 (F.P.M.d.V.)

ABSTRACT The laboratory mouse is the most widely used animal model for biomedical research, due in part to its well-annotated genome, wealth of genetic resources, and the ability to precisely manipulate its genome. Despite the importance of genetics for mouse research, genetic quality control (QC) is not standardized, in part due to the lack of cost-effective, informative, and robust platforms. Genotyping arrays are standard tools for mouse research and remain an attractive alternative even in the era of high-throughput whole-genome sequencing. Here, we describe the content and performance of a new iteration of the Mouse Universal Genotyping Array (MUGA), MiniMUGA, an array-based genetic QC platform with over 11,000 probes. In addition to robust discrimination between

most classical and wild-derived laboratory strains, MiniMUGA was designed to contain features not available in other platforms: (1) chromosomal sex determination, (2) discrimination between substrains from multiple commercial vendors, (3) diagnostic SNPs for popular laboratory strains, (4) detection of constructs used in genetically engineered mice, and (5) an easy-to-interpret report summarizing these results. In-depth annotation of all probes should facilitate custom analyses by individual researchers. To determine the performance of MiniMUGA, we genotyped 6899 samples from a wide variety of genetic backgrounds. The performance of MiniMUGA compares favorably with three previous iterations of the MUGA family of arrays, both in discrimination capabilities and robustness. We have generated publicly available consensus genotypes for 241 inbred strains including classical, wild-derived, and recombinant inbred lines. Here, we also report the detection of a substantial number of XO and XXY individuals across a variety of sample types, new markers that expand the utility of reduced complexity crosses to genetic backgrounds other than C57BL/6, and the robust detection of 17 genetic constructs. We provide preliminary evidence that the array can be used to identify both partial sex chromosome duplication and mosaicism, and that diagnostic SNPs can be used to determine how long inbred mice have been bred independently from the relevant main stock. We conclude that MiniMUGA is a valuable platform for genetic QC, and an important new tool to increase the rigor and reproducibility of mouse research.

The laboratory mouse is among the most popular and extensively used models for biomedical research. For example, in 2018 the word "mouse" appeared in the abstract of over 82,000 scientific manuscripts available in PubMed. The laboratory mouse is such an attractive model due to the existence of hundreds of inbred strains and outbred lines designed to address specific questions, as well as the ability to edit the mouse genome; originally by homologous recombination and now with more efficient and simple techniques such as clustered regularly interspaced short palindromic repeats (Lanigan *et al.* 2020). The centrality of genetics in mouse-enabled research begs the question of how genetic quality control (QC) is performed in these experiments. At a minimum, genetic QC should provide reliable information about the sex, genetic background, and presence of genetic constructs in a given sample in a robust and cost-effective manner.

We have a long track record of developing genotyping arrays for the laboratory mouse, from the Mouse Diversity Array (MDA, Yang *et al.* 2009) to the previous versions of the Mouse Universal Genotyping Array (MUGA) (Morgan and Welsh 2015). These tools were originally designed for the genetic characterization of two popular genetic reference populations, the Collaborative Cross (CC) and the Diversity Outbred, and then used in experiments involving other laboratory strains as well as wild mice (Yang *et al.* 2011; Collaborative Cross Consortium 2012; Carbonetto *et al.* 2014; Arends *et al.* 2016; Didion *et al.* 2016; Rosshart *et al.* 2017; Shorter *et al.* 2017; Srivastava *et al.* 2017; Veale *et al.* 2018).

In the following paragraphs, we discuss each of the main components of genetic QC as defined above. Sex is widely recognized as a key biological variable. Standard chromosomal sex determination in the MDA and MUGA relied on detection of the Y chromosome. This approach has limitations, most notably the inability to identify sex chromosome aneuploidies. In mammals, including humans and mice, sex chromosome abnormalities are relatively frequent (Searle and Jones 2002; Cheng *et al.* 2014), and thus the ability to detect them will substantially improve genetic QC.

The ability to discriminate between genetic backgrounds is critical for genetic QC, and all previous platforms were able to accomplish this goal to varying degrees. Array-based discrimination depends on the number of markers, their spatial distribution, and the ascertainment bias of those markers. While the MDA and several previous iterations of the MUGA had tens to hundreds of thousands of markers, the selection of those markers depended heavily on whole-genome sequencing (WGS) data from <20 inbred strains (Yang *et al.* 2007, 2009, 2011; Keane *et al.* 2011; Morgan *et al.* 2015). Thus, these platforms provided very fine-grained discrimination for some strains and coarse or happenstance discrimination for many others. An extreme example of the latter is the very poor discrimination between substrains. Mouse genetics is built on the phenotypic differences observed between strains and, recently, we have come to appreciate that closely related substrains can be phenotypically divergent due to variants accumulated by genetic drift (Kumar *et al.* 2013; Treger *et al.* 2019). Drift within an inbred strain can also lead to phenotypic divergence.

The presence of a genetic construct is designed to make carriers phenotypically different from noncarriers (Lanigan *et al.* 2020). Thus, the ability to detect genetic constructs will enhance genetic QC. Currently, detection of constructs is achieved by custom PCR designed for each construct in a given mouse stock. Because this process is costly and time-consuming, most researchers only test for the desired construct(s) used in their experiment. Many mouse stocks are the product of breeding mice with different constructs (*e.g.*, flox-flanked knockouts and Cre recombinase are

[1]These authors contributed equally to this work.
[2]Corresponding author: Department of Genetics, University of North Carolina, CB #7264, Genetic Medicine Building, Chapel Hill, NC 27599. E-mail: fernando_pardo-manuel@med.unc.edu

simultaneously present in many stocks). This cross-breeding and the common process of sharing genetically modified stocks between groups can lead to the accumulation of genetic constructs. A genetic QC platform that tests for multiple commonly used constructs will therefore be highly desirable. The MDA and the first two iterations of the MUGA arrays were not designed to detect any constructs (Yang *et al.* 2009; Morgan *et al.* 2015). Efforts to extend the use of the MUGA to detect genetic constructs were met with limited success in GigaMUGA (Morgan *et al.* 2015). We conclude that current genotyping tools are suboptimal for construct detection. For microarray-based construct detection, the most valuable assays are those that can detect the most popular constructs independent of site of insertion and the genetic background of the sample.

In addition to its value as a genetic QC tool, a well-designed genotyping array can also be a valuable tool for experimental research. Two such areas of research are sex chromosome biology and genetic mapping using reduced-complexity crosses (RCC; Kumar *et al.* 2013; Bryant *et al.* 2020). RCC are predicated on the idea that if a heritable phenotype is variable between a pair of closely related laboratory substrains, then QTL mapping combined with a complete catalog of the few thousand variants that differ among these substrains can lead to the rapid identification of the candidate causal variants (Kumar *et al.* 2013; Babbs *et al.* 2019). The exact number of variants between a pair of substrains, or within a set of substrains, varies substantially but is several orders of magnitude fewer than between classical strains (Mortazavi *et al.* 2020; M. T. Ferris, unpublished data). This addresses one of the major limitations of standard mouse crosses, namely the cost in time and resources to move from QTL to quantitative trait variants (Scalzo and Yokoyama 2008). The RCC concept has been successfully demonstrated in crosses between C57BL/6J and C57BL/6NJ (Kumar *et al.* 2013; Babbs *et al.* 2019), but existing genotyping platforms do not support extension to other crosses because of the lack of sufficient informative markers to support robust QTL mapping. Identification of such markers requires WGS from all parental substrains used in the RCC. By definition, most of these markers will be diagnostic for one substrain, thus improving genetic background identification.

To address these limitations, we created a fourth iteration of the MUGA family of arrays that we call MiniMUGA. The central considerations for the design were to reduce genotyping costs, robustly determine chromosomal sex, provide broad discrimination between most inbred strains and substrains, and reliably detect the presence of popular genetic constructs. We also incorporated diagnostic variants for multiple substrains to expand RCC to crosses between those substrains. MiniMUGA fulfills all our criteria and facilitates simple, uniform, and cost-effective standard genetic QC, as well as serving the mouse community at large by providing a new tool for genetic studies.

Materials and Methods

Reference samples

To test the performance of the MiniMUGA array, we genotyped 6899 DNA samples from a wide range of genetic backgrounds, ages, and tissues (Supplemental Material, Table S1). These samples include examples of inbred strains, F1 hybrids, experimental crosses, and cell lines (Table 1). The array content was designed in two phases, resulting in preliminary and production versions of the array. We genotyped 5604 samples in the preliminary version of the array containing 10,171 markers. We genotyped 1295 samples in the production version of the array. The production version of the array includes 954 additional markers selected to increase coverage of diagnostic SNPs for selected substrains (905 markers targeting 39 substrains) and additional constructs (45 markers targeting seven constructs). Samples were genotyped to determine the marker performance and information content, and to develop the multiple pipelines discussed throughout this paper. Overall, 6300 samples were genotyped once and 225 samples were genotyped two or more times, resulting in a total of 6525 unique samples genotyped.

Table S1 provides comprehensive information about each of these samples including name, type, whether it was genotyped in the preliminary or production version of the array, whether it was used in the array calibration process, and whether the sample was used to determine consensus genotypes or thresholds for chromosomal sex determination. Table S1 also lists chromosomal sex, basic QC metrics, and values used to determine the presence of 17 constructs. A complete description of the information provided in Table S1 is available in the table legend.

DNA stocks for inbred strains were purchased from The Jackson Laboratory over a decade ago, or provided by the authors. DNA from most other samples was prepared from tail clips or spleens using the DNeasy Blood & Tissue Kit (catalog no. 69506; QIAGEN, Valencia, CA). Approximately 1.5-μg genomic DNA per sample was shipped to Neogen Inc. (Lincoln, NE) for array hybridization and genotype calling.

Microarray platform and genotype calling

MiniMUGA is implemented on the Illumina Infinium XT platform (Illumina, Inc., San Diego, CA). Invariable oligonucleotide probes 50 bp in length are conjugated to silica beads that are then addressed to wells on a chip. Sample DNA is hybridized to the oligonucleotide probes and a single-base-pair templated extension reaction is performed with fluorescently labeled nucleotides (Steemers *et al.* 2006). The relative signal intensity from alternate fluorophores at the target nucleotide is processed into a discrete genotype call (AA, AB, or BB) using the Illumina GenomeStudio genotyping software (Illumina). Although the two-color Infinium readout is optimized for genotyping biallelic SNPs, both the total and relative signal intensity can also be informative for copy-number variation and construct detection. For each marker in the

preliminary array, we optimized the default clustering algorithm with a training data set of 2698 high-quality samples representing a wide variety of genetic backgrounds. Similarly, the production content markers were calibrated using 1295 samples (Table S1).

Probe design

Of the 11,125 markers present in the production version of the array, 10,819 (97.2%) are probes designed for biallelic SNPs and the remaining 306 markers (2.6%) are probes designed to test for the presence of genetic constructs (Table S2). Nucleotides are labeled such that only one silica bead is required to genotype most SNPs, except the cases of A/T and C/G SNPs, which require two beads. To maximize information content, target SNPs were biased toward single-bead SNPs (mostly transitions). There are 10,721 single-bead assays and 404 two-bead assays. All construct probes are single-bead assays. The transition:transversion ratio in SNPs (excluding constructs) is 3:1.

Probe annotation

Probe design and performance of individual assays was used to annotate the array. Table S2 contains the following information: (1) marker name, (2) chromosome, (3) position, (4) strand, (5–6) sequences for one- and two-bead probes, (7–8) reference and alternate alleles at the SNP, (9) tier, (10) reference SNP ID# (rsID), (11) diagnostic information, (12) uniqueness, (13) X chromosome markers used to determine the presence and number of X chromosomes, (14) Y chromosome markers used to determine the presence of a Y chromosome, and (15) markers added in the production version. A complete description of the information provided in Table S2 is available in the table legend.

Chromosomal sex determination

We selected a set of 2348 control samples (1108 males and 1240 females) with known X and Y chromosome number, as determined through standard anatomical sexing and/or known reproductive status. In the case of mice with known and well-defined genetics (inbreds and F1s), this was further confirmed by homozygous or heterozygous status at chromosome X markers. For each sample, we first normalized the intensity values at each X and Y chromosome marker by dividing the intensity (r) by the median intensity at all of the autosomal markers in that sample. These autosome-normalized intensity values are used in all subsequent sex-determination calculations.

The next step of chromosomal sex determination was to identify sex-linked markers that provide an estimate of sex chromosome number consistent with the anatomical sex and that have low between-sample noise. We identified 269 X and 72 Y sex-informative markers as those for which the ranges of median normalized intensity, as defined by their standard deviations not overlapping between male and female controls (Figure S1). The identity of these markers is provided in Table S2.

Next, we established chromosomal sex intensity threshold values. For each sample, we plotted the medians of the normalized intensity values at the X-informative markers on the x-axis and the medians of the normalized intensity values at the Y-informative markers on the y-axis (Figure 1). Based on this plot we identified two clusters, one containing the control males and one containing the control females (XY and XX, respectively). These two clusters contain >99% of the samples. Two additional clusters represent XO and XXY aneuploids, and are located at the predicted X and Y areas based on chromosomally normal males and females. Some XO and XYY cases are confirmed through genetic analysis (see the *Results*). We defined chromosomal sex (XX, XY, XO, and XXY) thresholds as the midpoint between the relevant clusters. There is a single Y threshold value (0.3) separating samples with or without a Y chromosome. We identified two independent X threshold values (0.77 and 0.69) depending on whether the sample has a Y chromosome or not (Figure 1). These threshold values were used to classify the chromosomal sex of experimental samples into four groups: XX, XY, XO, or XXY.

Generation of consensus genotypes

The impetus for creating consensus genotypes for inbred strains in MiniMUGA is to provide a set of reference genotype calls for widely used strains. When possible, we included both sexes and multiple biological and technical replicates of a given inbred strain to smooth over any errors in genotyping results, identify problematic markers, and to provide a more robust set of reference calls for comparison.

For each of 241 inbred and recombinant inbred strains (Table S3), we genotyped between 1 and 19 replicates (average 3.2 per strain). Most inbred strains (179) were genotyped more than once. For 53 strains (mostly recombinant inbred lines from the BXD panel) we did not genotype a male, and thus Y chromosome genotypes are not provided for those strains. Over one-half of the strains (146) were genotyped only in the preliminary version of the array, so genotypes at markers added to the production version of the array are missing in those strains. See Table S1 for details.

We generated consensus genotype calls at all 10,819 autosomal, X, pseudoautosomal region (note that this region may vary among strains) (Morgan *et al.* 2019), and Y chromosome markers (biallelic SNPs). For each strain, at each marker, we recorded the genotype calls in all of the constituent samples and determined the consistency among these calls. For strains with more than one sample, if all calls were consistent, the consensus genotype is shown in upper case (A, T, C, G, H, or N). We define partially consistent calls as those with a mix of one or more calls of a single nucleotide (A, T, C, or G), and one or more H and/or N calls. Partially consistent calls are shown in lower case, as are calls for strains with a single constituent sample. Inconsistent calls are those for which two distinct nucleotides calls are observed. Unless noted, inconsistent genotypes within a strain are shown as N in the consensus. Partially diagnostic SNPs (see below) are

always inconsistent (both allele calls are present in a single substrain), and the consensus call is the diagnostic allele shown in lower case. For CC strains, inconsistent consensus genotypes are shown as H, as these strains are known to be not fully inbred (Srivastava *et al.* 2017; Shorter *et al.* 2019). For mitochondria and *Y* chromosome markers, consensus calls follow the same rules except H calls are treated as N. Table S4 provides a list of rules for generating all possible consensus calls. Table S5 provides a listing of the consensus genotypes.

Informative SNPs between closely related substrains

To increase the specificity of MiniMUGA as a tool for discriminating between closely related inbred strains, we used public data from several other studies providing genotype or WGS information (Yang *et al.* 2009; Keane *et al.* 2011; Adams *et al.* 2015; Morgan *et al.* 2015). Most importantly, we included SNPs that are segregating between substrains. These SNPs were identified by WGS of 33 substrains (Table 2). These genome sequences will be made available as part of an upcoming publication (M. T. Ferris, R. S. Baric, M. T. Heise, C. M. Sassetti, and F. Pardo-Manuel de Villena, unpublished results). Finally, we included 339 variants discriminating between substrains of C57BL/10 (provided by A. A. Palmer, Y. Ren, and C. L. St. Pierre). The preliminary version of the array included 5171 probes present in GigaMUGA (Morgan *et al.* 2015). These were included to cover the genome uniformly in classical and wild-derived inbred strains, and a few were also informative between substrains.

Probes for genetically engineered constructs

We designed and included 306 probes targeting commonly used genetic constructs (257 in the preliminary phase and 49 in the production phase; Table S6). We identified conserved 51-mers that fulfill the following three conditions: (1) they are present in construct sequences available from either Addgene or GenBank, (2) the 5′ 50-mers did not have matches in the mouse genome, and (3) the last base pair of the 51-mer is an A in the forward orientation or a T in the reverse orientation. The alternative allele is either a C in the forward orientation or a G in the reverse orientation. This alternative allele is a requirement for Illumina array design and does not represent a true SNP. Genotype calls at construct probes are not relevant.

As these probes are only useful in the context of intensity, and not in genotype calls, we developed a different pipeline to classify these probes into tiers (Morgan *et al.* 2015; applied in the *Results* to genomic SNPs). Because of the probe design, we only assessed probes for their ability to have consistent low raw normalized intensity in the *x*-axis in samples with nonmanipulated genomes (585 samples from inbred strains and 250 F1 hybrids, referred to as negative construct controls in Table S1) and at least some samples with high raw normalized intensity in our experimental data set. We eliminated 79 probes from the analysis because they failed this step (probes with purple labels in Figure S2A). We further eliminated 50 probes from the analysis because the range of variation in our experimental samples was not sufficiently distinguishable from the negative controls, or we observed a unimodal intensity distribution in the experimental samples (probes with blue labels in Figure S2A). Our experience with GigaMUGA suggested that a single construct probe was insufficient to robustly classify samples with the presence or absence of a construct (Morgan *et al.* 2015). Therefore, we eliminated 14 probes from the analysis because of low correlation of intensities across our experimental sample set (probes with red labels in Figure S2, A and B). Figure S2B shows the clusters based on the correlation of probe intensities. Finally, we confirmed that clustered probes were targeting the same or related constructs based on the Basic Local Alignment Search Tool (Boratyn *et al.* 2013). These alignments are provided in Figure S3. In total, these 163 probes mapped to 17 biologically distinct constructs (see Table 3). For each of these constructs, we identified conservative threshold values for the presence and absence based on the sum intensity of the probes assigned to that construct. We used the distribution of values to identify breaks and set the thresholds such that we minimized the number of samples misclassified as positive or negative. Positive controls (when available) were used to validate our classification schema.

Additional sample quality metrics

Most quality metrics for genotyping arrays are based on genotype calls. However, intensity-based analyses, such as chromosomal sex determination, assume quasi-normal distribution of marker intensities in a given sample (Figure S4). In our data set, some samples had significantly skewed and idiosyncratic intensity distributions. Among these samples, there were several erroneously identified as sex chromosome aneuploids.

To identify samples with abnormal intensity distributions, we used 200 random samples with no chromosomal abnormalities and confirmed that, in aggregate, they have quasi-normal intensity distribution (reference distribution) at the autosomal markers of the preliminary array content. We then computed a power divergence statistic (pd_stat; equivalent to Pearson's chi square goodness of fit statistic) for each sample, comparing its autosomal intensity distribution to that of the reference distribution. Figure S5 shows the distribution of pd_stat values in our entire data set. We selected a pd_stat value of 3230 as the threshold, and in samples with higher values the reported chromosomal sex could be incorrect. The pd_stat should be carefully scrutinized for samples with reported sex chromosome aneuploidy. The threshold also ensures that in samples from species other than *Mus musculus*, chromosomal sex determination is treated with skepticism.

To determine whether a high pd_stat had an effect on the accuracy of genotyping calls we selected four pairs of different F1 mice [(A/JxCAST/EiJ)F1_M15765; (CAST/EiJxA/J) F1_F002; (CAST/EiJxNZO/HlLtJ)F1_F0019; (CAST/EiJxNZO/ HlLtJ)F1_F022; (NZO/HlLtJxNOD/ShiLtJ)F1_F0042; (NZO/ HlLtJxNOD/ShiLtJ)F1_F0042; (PWK/PhJxNZO/HlLtJ)

F1_F0019 and (PWK/PhJxNZO/HlLtJ)F1_M0001] that cover a variety of pd_stat contrasts (high/low, medium/medium, and low/low). For each pair, we first determined the pairwise consistency of the genotype calls and then compared these genotypes to predicted calls for the consensus reference parental inbred strains. Pairwise comparison consistencies in the autosomes (excluding N calls) vary between 99.5 and 100%. Similarly, the consistency with predicted genotypes is very high (99.5–100%). We conclude that the pd_stat is independent of genotype call quality.

Data availability

Genotype calls and hybridization intensity data (both raw and processed) for 6899 samples, and consensus genotypes for 241 inbred strains are available for download at the University of North Carolina at Chapel Hill (UNC) Dataverse. These data are posted at https://dataverse.unc.edu/dataverse/MiniMUGA. Supplemental material available at figshare: https://doi.org/10.25386/genetics.11971941.

Results

Sample set, reproducibility, and array annotation

The 599 technical replicates (Table S1, see *Materials and Methods*) were used to calculate the reproducibility of the genotype calls. Overall, 99.6 ± 0.4% of SNP genotype calls were consistent between technical replicates (range 95.9–100%). The consistency rate was similar for replicates run in the same and different versions of the array. Samples with lower consistency rates included samples from more distant species and subspecies (SPRET/EiJ, SFM, SMZ, MSM/MsJ and JF1/Ms), lower-quality samples, and cell lines. Inconsistency was typically driven by a small minority of markers and by "no calls" in one or few of the technical replicates.

We annotated the array based on probe design and performance of individual assays was used to annotate the array (Table S2). Probes were classified in four tiers based on the presence of reference, alternate, and H calls in our sample set. Probes in tier 1 and 2 (all three genotype calls present and two genotype calls present, respectively) were used in most of the genotype-based analyses. For tier 3 probes, only one genotype call was present, while for tier 4 all samples had N calls. For construct markers, this tier classification was not relevant.

Improved chromosomal sex determination reveals sex chromosome aneuploidy due to strain-dependent paternal nondisjunction

Typically, genetic determination of sex of a mouse sample has relied on detecting the presence of a Y chromosome. This approach does not estimate X chromosome dosage and thus lacks the ability to identify samples with common types of sex chromosome aneuploidies. MiniMUGA uses probe intensity to discriminate between normal chromosomal sexes (XX and XY) and two types of sex chromosome aneuploidies, XO and XXY (Table S1). Our methodology (*Materials and Methods*)

provides a robust framework to discriminate between at least four types of chromosomal sex (Figure 1). Our set of 6899 samples was composed of 3507 unique females (no Y chromosome present) and 3018 unique males (Y chromosome present).

We initially identified 54 samples as potential XO and XXY. However, in eight XO females the pattern of heterozygosity and recombination in the X chromosome (Table S7) demonstrated that these were, in fact, normal XX females with abnormal intensity distributions and pd_stat values above the threshold (see *Materials and Methods*). Once these eight samples were removed, 46 samples that had sex chromosome aneuploidies remained. To determine the rate of aneuploidy we only considered unique samples (not replicates). This resulted in 45 aneuploid samples among 6525 total unique samples, an overall 0.7% rate. This rate was driven by a highly significant excess ($7X$, $\chi^2 = 62.9384$; $P < 0.00001$) of sex chromosome aneuploids among the cell lines. Notably, all these aneuploids were XO. Among live mice there were 36 unique aneuploids (a rate of 0.55%). This rate was similar to but higher than that previously reported in both mice and in humans (Searle and Jones 2002; Chesler *et al.* 2016; Le Gall *et al.* 2017). In this data set, unique XO females were observed at significantly higher frequency than unique XXY males ($P = 0.02$; 25 XO females and 11 XXY males) (Table 1).

For 22 of the 45 unique samples with sex chromosome aneuploidies, the parents were known and had informative markers in the X chromosome. This information allowed us to potentially determine the parental origin of the missing (in XO) or the extra (in XXY) X chromosome based on the parental haplotype inherited and the presence of recombination in the X chromosome (Figure 2 and Table S7). Overall, the parental origin can be determined unambiguously in 21 of these samples, and in all but one sample (95%) the aneuploidy is due to sex chromosome nondisjunction in the paternal germ line (Figure 2). Note that this applies to both XO and XXY samples. Given the paternal origin of most sex chromosome aneuploidies, we investigated whether the type of sire had an effect on this phenomenon. We observed a significantly ($P < 0.00001$) higher rate of aneuploids in the progeny of (CC029/Unc × CC030/GeniUnc) F1 hybrid males compared with all other types of sires. Out of 180 progeny of this cross, 5% of genotyped samples were aneuploids and both XO and XXY were observed (three XO and six XXY mice, respectively). There was also some evidence of a higher rate of sex chromosome aneuploids in progeny of sires with CC011/Unc background (five XO females, Table S7). We conclude that sex chromosome aneuploidy is relatively common in laboratory mice, originates predominantly in the paternal germ line, and depends on the sire genotype. In some backgrounds, aneuploidy rate is an order of magnitude higher than in the general population.

Detection of Y chromosome mosaicism

There were eight samples (two classified as XX, three as XXY, and three as XO) with abnormal chromosome Y intensities

Table 1 Sample set

Content	Chromosomal sex	Inbred	F1	CC	Cross	Unclassified	Cell lines	Total
Preliminary	XX	138	131	305	1383	817	87	2861
	XY	265	41	181	1236	907	74	2704
	XO	0	1	3	11	8	9	32
	XXY	0	1	1	2	3	0	7
Subtotal								5604
Production	XX	41	59	40	580	21	4	745
	XY	153	13	7	248	112	10	543
	XO	0	1	0	2	0	0	3
	XXY	0	0	0	4	0	0	4
Subtotal								1295
Total		597	247	537	3466	1868	184	6899

The table provides the number of samples genotyped in the preliminary and production version of the array classified according their chromosomal sex and type. CC, collaborative cross; Cross, experimental back- and intercrosses; unclassified, samples provided by the coauthors that may be of any type.

(either too low or too high) and with low number of chromosome Y genotype calls [between 6 and 56 calls compared with 62.99 (mean) ± 1.3 (SD) across all males]. These eight samples are standard laboratory mice and are shown in gray in Figure 1. The intensity and genotypes strongly suggest the presence of a Y chromosome that can be explained either by mosaicism or an incomplete Y chromosome. We performed several additional analyses to discriminate between these two explanations.

As a test case, we selected the tail-derived sample TL9348 (Tables S1 and S8) because it was expected to be an F1 hybrid male derived from a C57BL/6J and 129X1/SvJ outcross (Figure 3A), and chromosomal sex was not questionable based on its pd_stat (137). Based on chromosome intensity, this sample was classified as an XXY male with low chromosome Y intensity (Figure 3B). Inspection of the genotype calls on chromosome X revealed a significant excess of N calls compared with the autosomes ($P < 0.00001$, Table S8). Furthermore, the H calls on the X chromosome in this sample only occurred at SNPs where C57BL/6J and 129X1/SvJ have different alleles, but these H calls occur only at a fraction of expected sites. These H calls are present over the entire X chromosome. The fact that sample TL9348 has approximately one half of the Y chromosome intensity of XY or XXY males, and that there is evidence of heterozygosity on the X chromosome, suggests that the mosaicism is due to the loss of both the Y chromosome and one of the two X chromosomes in a fraction of cells. To test this hypothesis, we plotted the intensity of informative X chromosome markers for three types of controls—C57BL/6J, 129X1/SvJ, and F1 hybrid females—derived from those two inbred strains, as well as for the suspected mosaic sample TL9348 (Figure 3C). For sample TL9348, the vast majority of the informative markers were clustered between the C57BL/6J and the F1 hybrid genotypes (Figure 3C). This pattern explains the observed mix of N calls, heterozygous calls, and C57BL/6J calls in sample TL9348 and confirms its mosaic nature. It further demonstrates that the X chromosome lost is the one from 129X1/SvJ. Based on the positions of the intensities for the Y and X chromosome makers in Figure 3, A and B, respectively, we

concluded that slightly less cells were XXY than XO (Figure 3, C and D). Considered together, these results indicate that the embryo started as an XXY due to paternal nondisjunction of the sex chromosomes and that mosaicism occurred in early development, a common observation in embryo mosaicism in humans (Johnson et al. 2010; Fragouli et al. 2011; McCoy 2017).

Among the remaining seven potential mosaics, one was a cell line and thus mosaicism of the sex chromosomes was not unexpected. For the other six samples we performed a similar analysis as the one described above. In all cases, the Y chromosome calls were consistent with those expected from their sires. This is consistent with Y chromosome mosaicism and not with sample contamination. However, only the two samples with 50 or more genotype calls on the Y chromosome have strong support for such a conclusion. In the Discussion we expand this analysis and provide some guidance for users of the array.

Strain-specific chromosome Y duplications

Among XY males there was a distinct cluster of 64 male samples with higher normalized median Y chromosome intensity (Figure 1). These samples include five inbred C3H/HeJ, two F1 hybrid males with a C3H/HeJ chromosome Y (Figure 4A), and 52 males derived from a C3H/HeJ by C3H/HeNTac F2 intercross. The plot of the normalized Y chromosome intensities in these males and 81 additional males with Y chromosomes derived from other C3H/He substrains (Figure 4A) revealed a clear separation between males carrying a Y chromosome from C3H/HeJ and males carrying C3H/HeNCrl, C3H/HeNHsd, C3H/HeNRj, C3H/HeNTac, and C3H/HeOuJ Y chromosomes. Males with the high-intensity Y chromosome also include two transgenic strains from The Jackson Laboratory: B6C3-Tg(APPswe, PSEN1dE9)85Dbo/Mmjax and B6;C3-Tg(Prnp-SNCA*A53T)83Vle/J. Both strains were developed and/or maintained on a B6C3H background (JAX Stocks 34829-JAX and 004479, respectively).

To determine the origin of the higher median intensity in males with a C3H/HeJ Y chromosome, we plotted the normalized intensities at 59 MiniMUGA markers located on the
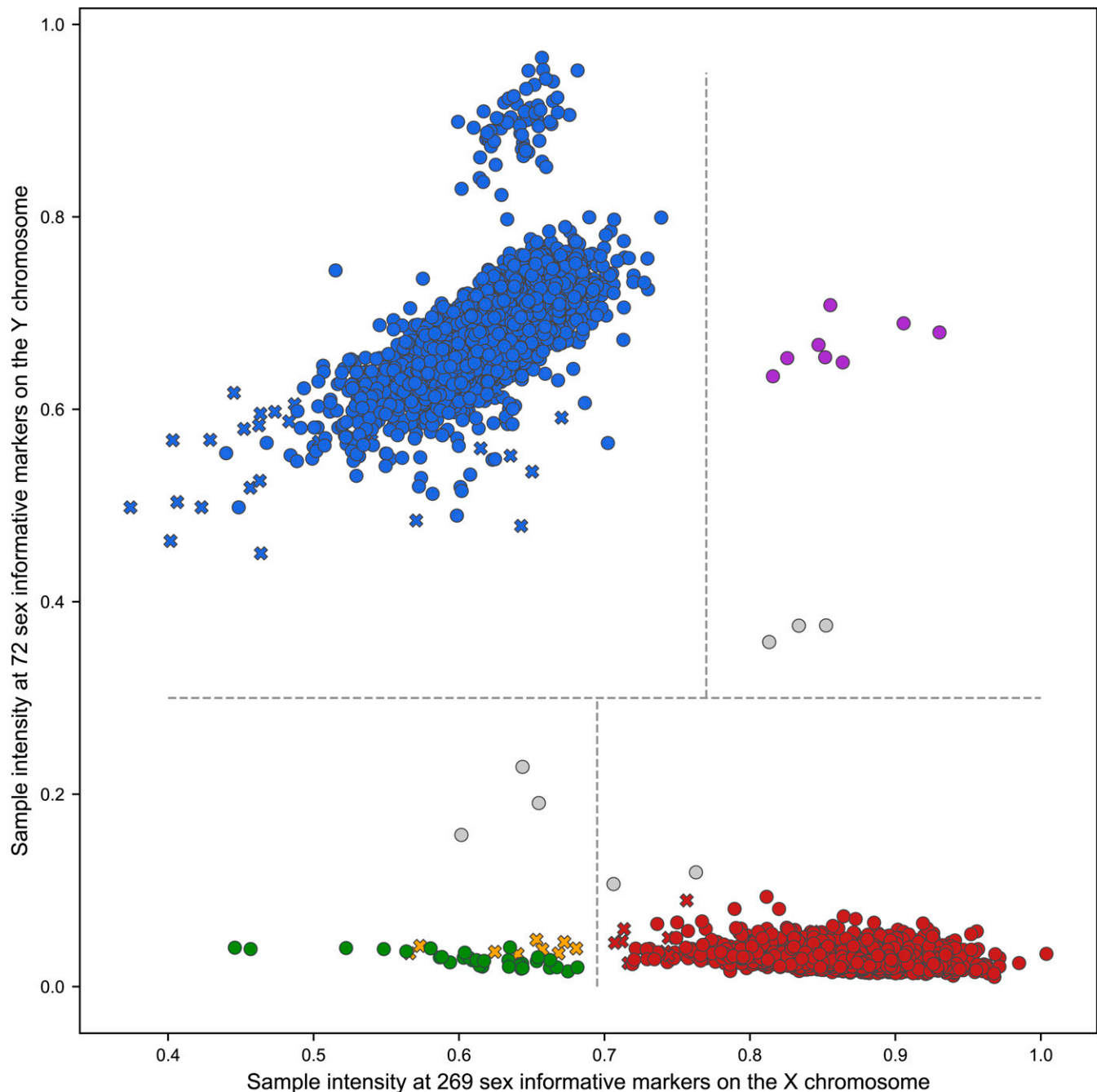
Figure 1 Chromosomal sex determination in 6899 samples. Each circle and cross represent one genotyped sample. The *x*-axis value is the autosome-normalized median sample intensity at 269 sex-informative *X* chromosome markers, and the *y*-axis value is the autosome-normalized median sample intensity at 72 sex-informative *Y* chromosome markers. The dot color denotes the assigned chromosomal sex: *XX*, red; *XY*, blue; *X*O, green; and *XXY*, purple. Potential mosaic samples are shown in gray and known errors in yellow. Samples with pd_stat lower than the threshold are shown as circles and samples with high pd_stat are shown as crosses.

short arm of chromosome *Y* and the four most proximal markers on the long arm of that chromosome (Figure 4B). Inspection of this figure indicates that 54 consecutive markers have distinctly higher intensities in C3H/HeJ males and are flanked by markers with intensities that are undistinguishable from males with other C3H/He *Y* chromosomes. These markers define a 2.9-Mb region located on the short

arm of the *Y* chromosome containing seven known genes—*Eif2s3y, Uty, Ddx3y, Usp9y, Zfy2, Sry*, and *Rbmy*—and 12 gene models (Figure 4B). We conclude that C3H/He substrain differences are due to an intrachromosomal duplication that arose and was fixed in the C3H/HeJ lineage after the isolation of that substrain in 1952 (Akeson *et al.* 2006). There are five additional non-C3H/He samples with

Table 2 Sequenced inbred mouse strains used to select the content of the genotyping array.

Background	Strain group	Diagnostic type	Full	Partial	Reference
129P2/OlaHsd	129P	Substrain	25	0	Keane *et al.* (2011); Doran *et al.* (2016)
129P3/J	129P	Substrain	54	0	M. T. Ferris *et al.*, unpublished results
129S1/SvImJ	129S	Substrain	82	13	Keane *et al.* (2011); Doran *et al.* (2016)
129S2/SvHsd	129S	Substrain	7	1	M. T. Ferris *et al.*, unpublished results
129S2/SvPasOrlRj	129S	Substrain	36	0	M. T. Ferris *et al.*, unpublished results
129S4/SvJaeJ	129S	Substrain	45	0	M. T. Ferris *et al.*, unpublished results
129S5/SvEvBrd	129S	Substrain	12	0	Keane *et al.* (2011); Doran *et al.* (2016)
129S6/SvEvTac	129S	Substrain	41	0	M. T. Ferris *et al.*, unpublished results
129T2/SvEmsJ	129T	Substrain	38	0	M. T. Ferris *et al.*, unpublished results
129X1/SvJ	129X	Substrain	39	0	M. T. Ferris *et al.*, unpublished results
A/J	A	Substrain	58	7	Keane *et al.* (2011); Doran *et al.* (2016)
A/JCr	A	Substrain	53	0	M. T. Ferris *et al.*, unpublished results
A/JOlaHsd	A	Substrain	38	0	M. T. Ferris *et al.*, unpublished results
BALB/cAnNCrl	BALB/c	Substrain	36	2	M. T. Ferris *et al.*, unpublished results
BALB/cAnNHsd	BALB/c	Substrain	109	4	M. T. Ferris *et al.*, unpublished results
BALB/cByJ	BALB/c	Substrain	3	4	M. T. Ferris *et al.*, unpublished results
BALB/cByJRj	BALB/c	Substrain	19	0	M. T. Ferris *et al.*, unpublished results
BALB/cJ	BALB/c	Substrain	103	3	Keane *et al.* (2011); Doran *et al.* (2016)
BALB/cJBomTac	BALB/c	Substrain	47	0	M. T. Ferris *et al.*, unpublished results
C3H/HeJ	C3H/He	Substrain	166	2	Keane *et al.* (2011); Doran *et al.* (2016)
C3H/HeNCrl	C3H/He	Substrain	39	0	M. T. Ferris *et al.*, unpublished results
C3H/HeNHsd	C3H/He	Substrain	39	1	M. T. Ferris *et al.*, unpublished results
C3H/HeNRj	C3H/He	Substrain	42	0	M. T. Ferris *et al.*, unpublished results
C3H/HeNTac	C3H/He	Substrain	45	14	M. T. Ferris *et al.*, unpublished results
C57BL/6J	C57BL/6	Substrain	136	20	Sarsani *et al.* (2019)
C57BL/6JBomTac	C57BL/6	Substrain	41	2	M. T. Ferris *et al.*, unpublished results
C57BL/6JOlaHsd	C57BL/6	Substrain	43	0	M. T. Ferris *et al.*, unpublished results
C57BL/6NJ	C57BL/6	Substrain	37	7	Keane *et al.* (2011); Doran *et al.* (2016)
C57BL/6NRj	C57BL/6	Substrain	20	0	M. T. Ferris *et al.*, unpublished results
B6N-Tyr < c-Brd>/BrdCrCrl	C57BL/6	Substrain	21	10	M. T. Ferris *et al.*, unpublished results
DBA/1J	DBA/1	Substrain	70	0	Keane *et al.* (2011); Doran *et al.* (2016)
DBA/1LacJ	DBA/1	Substrain	77	2	M. T. Ferris *et al.*, unpublished results
DBA/1OlaHsd	DBA/2	Substrain	32	0	M. T. Ferris *et al.*, unpublished results
DBA/2J	DBA/2	Substrain	112	0	Keane *et al.* (2011); Doran *et al.* (2016)
DBA/2JOlaHsd	DBA/2	Substrain	39	0	M. T. Ferris *et al.*, unpublished results
DBA/2JRj	DBA/2	Substrain	30	0	M. T. Ferris *et al.*, unpublished results
DBA/2NCrl	DBA/2	Substrain	85	14	M. T. Ferris *et al.*, unpublished results
DBA/2NTac	DBA/2	Substrain	36	10	M. T. Ferris *et al.*, unpublished results
FVB/NCrl	FVB	Substrain	47	0	M. T. Ferris *et al.*, unpublished results
FVB/NHsd	FVB	Substrain	39	1	M. T. Ferris *et al.*, unpublished results
FVB/NJ	FVB	Substrain	72	7	Keane *et al.* (2011); Doran *et al.* (2016)
FVB/NRj	FVB	Substrain	47	0	M. T. Ferris *et al.*, unpublished results
FVB/NTac	FVB	Substrain	37	0	M. T. Ferris *et al.*, unpublished results
NOD/MrkTac	NOD	Substrain	33	0	M. T. Ferris *et al.*, unpublished results
NOD/ShiLtJ	NOD	Substrain	51	3	Keane *et al.* (2011); Doran *et al.* (2016)
Subtotal			2281	127	
129S	129S	Strain group	17	0	
A	A	Strain group	57	0	
BALB/c	BALB/c	Strain group	125	0	
C3H/He	C3H/He	Strain group	45	0	
C57BL/10	C57BL/10	Strain group	291	0	Mortazavi *et al.* 2020
C57BL/6	C57BL/6	Strain group	19	0	
DBA/1	DBA/1	Strain group	5	0	
DBA/2	DBA/2	Strain group	62	0	
FVB/N	FVB/N	Strain group	2	0	
NZO	NZO	Strain group	12	0	Keane *et al.* (2011); Doran *et al.* (2016)
Subtotal			635	0	
Total			2916	127	

The table provides the strain name and group, the number and type for both fully and partial diagnostic SNPs, and the source of the whole-genome sequencing data.

Table 3 Validated constructs

Name	Abreviation	Number of probes	Number of distinct probes
"Greenish" Fluorescent Protein (EGFP, EYFP, and ECFP)	g_FP	19	19
SV40 large T antigen	SV40	18	18
Cre recombinase	Cre	16	12
Tetracycline repressor protein	tTA	14	14
Diptheria toxin	DTA	11	11
Human CMV enhancer *version b*	hCMV_b	10	7
Luciferase and firefly luciferase	Luc	10	10
Chloramphenicol acetyltransferase	chloR	9	9
Bovine growth hormone poly A signal sequence	bpA	8	4
iCre recombinase	iCre	8	8
Reverse improved tetracycline-controlled transactivator	rtTA	8	4
CRISPR associated protein 9	cas9	7	7
Blasticidin resistance	BlastR	6	4
Internal Ribosome Entry Site	IRES	6	6
hCMV enhancer *version a*	hCMV_a	5	4
"Reddish" fluorescent protein (tdTomato, mCherry)	r_FP	6	6
Herpesvirus TK promoter	hTK_pr	2	2
Total		163	145

The table lists the name, abbreviation shown in the report and the number of total and distinct probes for 17 constructs validated in the data set reported here. EGFP, enhanced green fluorescent protein; EYFP, enhanced yellow fluorescent protein; ECFP, enhanced cyan fluorescent protein; CMV, cytomegalovirus; hCMV, human cytomegalovirus; SV40, simian virus 40; TK, thymidine kinase.

high normalized median chromosome *Y* intensity, four technical replicates from a single DBA/1OlaHsd male and a single $Axl^{-/-}$ congenic mouse on a C57BL/6 background (Figure 4A). Each case represents an independent (different haplotype and different boundaries, Figure S6) and very recent duplication of the *Y* chromosome. These duplications were segregating within a closed colony. Given that we identified three independent large segmental duplications in the short arm of the *Y* chromosome among 3018 unique males (Table 1), this leads to a crude mutation rate estimate of 1/1000. This mutation rate is slightly lower than some segmental duplications in the mouse (Egan *et al.* 2007), higher than the mutation rate in microsatellites (Dallas 1992), and consistent with high levels of structural variation in the short arm of the *Y* chromosome in wild mice (Morgan and Pardo-Manuel de Villena 2017).

An effective tool for genetic QC in laboratory inbred strains

To determine the performance of MiniMUGA among inbred strains we genotyped 779 samples representing 241 inbred strains including 86 classical inbred strains, 34 wild-derived inbred strains, 49 BXD recombinant inbred lines, and 72 CC strains (Table S3). We created consensus genotypes for each inbred strain using both biological and technical replicates (see the *Materials and Methods*). The use of replicates strengthens genetic analyses as they provide a simple but robust method to determine the performance of each SNP in each strain (see the *Discussion*), as well as determining the dates when diagnostic alleles arise and potentially became fixed (see *Diagnostic SNPs as a tool for genetic QC and strain dating*). We note that for the CC strains, which are incompletely inbred (Srivastava *et al.* 2017; Shorter *et al.*

2019), our consensus calls should be treated with caution and viewed as preliminary. This is particularly true given that they were based on a small number of individuals sampled from the UNC Systems Genetics Core Facility colony (Morgan *et al.* 2015, http://csbio.unc.edu/CCstatus/index.py) between 2016 and 2017 (Shorter *et al.* 2019). Future sampling of a wider range of individuals from CC strains throughout the history of the CC colony will result in more accurate consensus genotypes for these strains.

Using the consensus genotypes we determined the number of informative markers for pairwise combinations of all inbred strains (excluding BXD and CC). Figure 5 summarizes the results for 83 classical inbred strains. Over 90% of comparisons have ≥1280 informative autosomal markers and all but 0.52% of pairwise comparisons have >40 informative autosomal markers (2.1 markers per autosome). These statistics are exceptional given the small number of markers in the array, and considering the number of diagnostic markers included and a substantial number of construct markers. Although our focus is on classical inbred strains, we extended the analysis to include 37 wild-derived strains. For all 2924 combinations of classical and wild-derived strains, the informativeness is high (mean = 3224, minimum = 1649, and maximum = 3827, data not shown). In marked contrast, combinations between wild-derived strains have a much wider range of informative SNPs (from 93 to 3410, data not shown) due to a fraction of combinations with few to a moderate number of informative SNPs. The pairs of strains with the lowest number of informative SNPs include pairs of strains from a taxa other than *M. musculus* (for example SPRET/EiJ, SMZ, and XBS) and pairs of strains that are known to have close phylogenetic relationships (TIRANO/EiJ and ZALENDE/EiJ; and PWD and PWK/PhJ) (Yang

Figure 2 Sex chromosome aneuploidy is due to paternal nondisjunction. The figure shows the parental sex chromosome and mitochondrial complement of the dam and sire for two types of crosses. Only the sex chromosomes and the mitochondria are shown. The *X* chromosomes are shown as long acrocentric, the *Y* chromosomes as shorter submetacentric, and the mitochondria as circles. The figure also shows the inferred parental origin of the sex chromosome aneuploidy and the actual number of cases observed in our data set. The sex chromosome configuration of standard types of sex chromosome aneuploidy in the progeny in each type of cross are shown with the inferred parental origin of the *X* chromosomes.

et al. 2011). We conclude that MiniMUGA improves cost-effective genotyping for dozens of standard laboratory strains and experimental crosses derived from them.

Mitochondria

MiniMUGA has 88 markers that track the mitochondrial genome, 82 of which segregate in our set of 241 inbred strains. Based on these 82 markers, the inbred strains can be classified into 22 different haplogroups, 19 of which discriminate between *M. musculus* strains (Figure 6A). Fifteen haplogroups represent *M. m. domesticus* (groups 1 to 15 in Figure 6A) two haplogroups represent *M. m. musculus* (16 and 17), and two *M. m. castaneus* (18 and 19). Three haplogroups represent different species such as *M. spretus* and *M. macedonicus*.

In *M. musculus*, nine haplogroups are present in multiple inbred strains while 10 are found in a single inbred strain. The most common haplogroup is present in 158 inbred strains (including 49 BXD and 26 CC strains). This haplogroup is found in many classical inbred strains including C57BL6/J, BALB/cJ, A/J, C3H/HeJ, DBA/1J, DBA/2J, and FVB/NJ. Unique haplogroups represent an interesting mix of wild-derived strains (LEWES/EiJ, CALB/Rk, WMP/Pas, SF/CamEiJ, TIRANO/EiJ, ZALENDE/EiJ, and CIM) and DBA/2 substrains (DBA/2JOlaHsd and DBA/2NCrl). CC strains fall into six common haplogroups, one shared by three CC founders (A/J, C57BL/6J, and NOD/ShiLtJ) and five haplogroups present in a single CC founder: PWK/PhJ,

129S1/SvImJ, CAST/EiJ, NZO/HlLtJ, and WSB/EiJ. Interestingly, SMZ, a wild-derived inbred strain of *M. spretus* origin, has a mitochondrial haplogroup that unambiguously clusters with *M. m. domesticus* (Figure 6A) demonstrating a case of interspecific introgression (Didion and Pardo-Manuel de Villena 2013).

Chromosome Y

MiniMUGA has 75 markers that track the *Y* chromosome, 57 of which segregate in our set of 189 inbred strains with at least one male genotyped. Based on these 57 markers, the inbred strains can be classified into 18 different haplogroups, 16 of which are *M. musculus* (Figure 6B). Only four haplogroups represent *M. m. domesticus*, two haplogroups represent *M. m. castaneus*, and 11 represent *M. m. musculus*. *M. spretus* and *M. macedonicus* are represented by a single haplogroup each. In *M. musculus*, all but one haplogroup (CIM) are present in multiple inbred strains. No single haplogroup dominates in our collection of inbred strains (the most common is present in 38 inbred strains). Interestingly, C57BL/6 substrains fall into three distinct haplogroups. The ancestral haplogroup is found in C57BL/6ByJ, C57BL/6NCrl, C57BL/6NHsd, C57BL/6NJ, C57BL/6NR,j and B6N-Tyr < c-Brd>/BrdCrCrl. This haplogroup is present in other classical inbred strains such as BALB/c, C57BL/10, C57BLKS/J, C57L/J, and C58/J. The second haplogroup is present in C57BL/6JBomTac, C57BL/6JEiJ, and C57BL/6JOlaHsd. Finally,
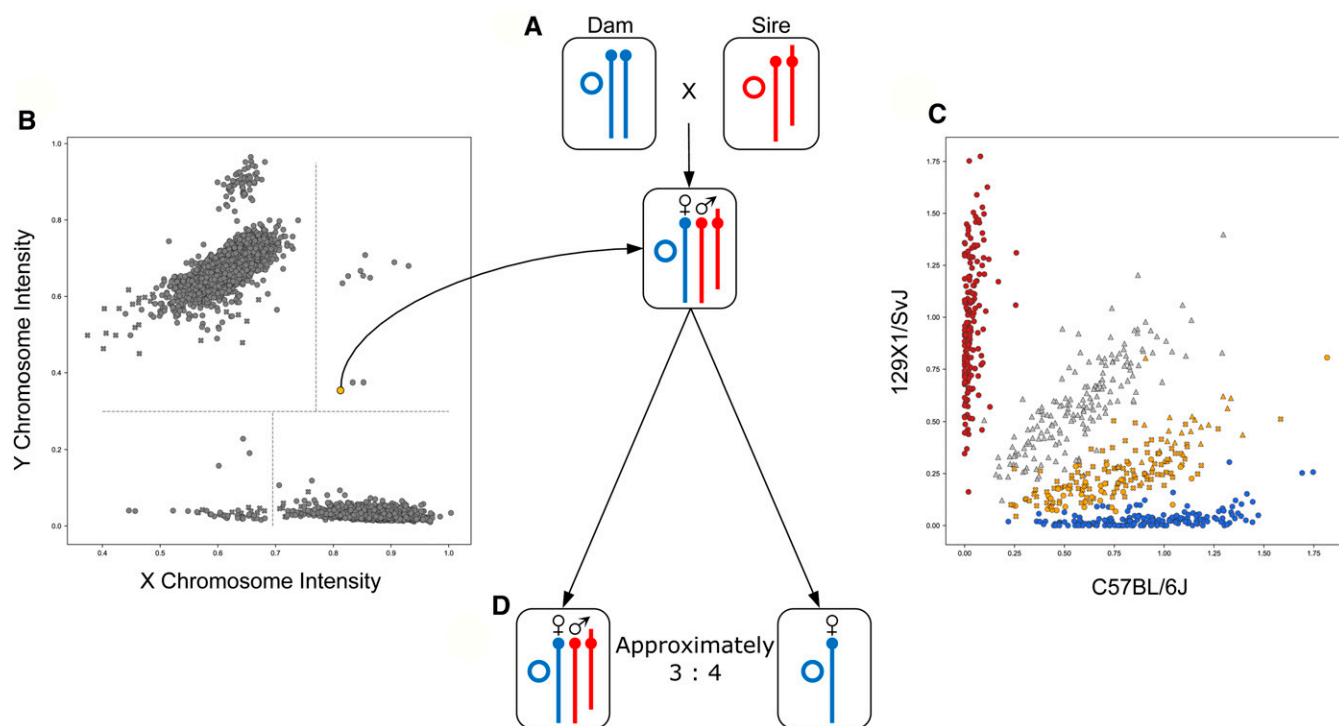
Figure 3 Complex sex chromosome aneuploidy and mosaicism in an F1 male. (A) The panel shows the chromosomal sex and mitochondria complement of the parents and F1 individual. Blue denotes C57BL/6J and red denotes 129X1/SvJ. (B) This panel is a reprint of Figure 1 and was used to classify the F1 male, shown as a yellow circle, as an *XXY* based on the x- and y-axis intensities (two *X* chromosomes and a *Y* chromosome present). This panel also provides evidence of mosaicism for the presence and absence of the *Y* chromosome (based on the low *Y* chromosome intensity). (C) This panel provides evidence of mosaicism for the *X* chromosome and identifies the paternal origin (129X1/SvJ) of the chromosome lost in some cells. The plot presents the intensities of the two alternate alleles for 173 *X* chromosome markers that are informative between the two parents. Four individuals are shown: a C57BL/6J female in blue, a 129X1/SvJ male in red, a (C57BL/6Jx129X1/SvJ)F1 female in gray, and the F1 male case in yellow. The shapes denote the type of call made by the Illumina software: circles are homozygous A, T, C, or G calls; triangles are H calls; and squares are N calls. (D) This panel shows the proposed sex chromosome complement of the two types of cells present in this F1 male case. This solution explains the observations from previous three panels.

C57BL/6J has its own private derived haplogroup shared with 10 CC strains. Each one of the eight founder strains of the CC (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HlLtJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ) has its own distinct haplogroup.

Diagnostic SNPs as a tool for genetic QC and sample dating

We define SNPs as diagnostic when the minor allele is present only in a single substrain or in a set of closely related substrains. The identification of these SNPs for inclusion in the array is based on WGS of 12 publicly available strains (Keane *et al.* 2011; Adams *et al.* 2015), 33 substrains sequenced by us, and SNP data for the C57BL/10 strain group (Table 2). Almost 30% of the SNPs (3045) in MiniMUGA are diagnostic. Although diagnostic SNPs have low information content (*i.e.*, most samples in a large set of genetically diverse mice will be homozygous for the major allele) they fulfill two critical objectives. First, they increase the specificity of the MiniMUGA array to identify the genetic background present in a sample. In addition, they are essential to extend the power of genetic mapping in RCC beyond the C57BL/6J-C57BL/6NJ paradigm (Kumar *et al.* 2013; Treger *et al.* 2019).

The 3045 diagnostic SNPs can be divided into two classes based on whether they are diagnostic for a specific substrain (*i.e.*, BALB/cJBomTac or C3H/HeJ) or two or more substrains within a strain group (*i.e.*, BALB/c or C3H/He). There are 2408 SNPs that are diagnostic for one of 45 substrains and 637 SNPs diagnostic for one of 10 strain groups (Table 2). A second classification divides diagnostic SNPs into fully diagnostic (2910) and partially diagnostic SNPs (129). The difference between these two classes is based on whether the diagnostic allele was fixed or was still segregating in the samples used to determine the consensus genotypes of 45 classical inbred strains.

All diagnostic SNPs originated as partially diagnostic SNPs and they highlight the often-overlooked fact that mutations arise in all stocks and some become fixed despite the best efforts to reduce their frequency and impact (Sarsani *et al.* 2019). Note that the classification of a SNP as partially diagnostic depends on the samples used for the consensus calls.

It is theoretically possible to date when diagnostic SNPs arose and whether and when they became fixed in the main stock of a substrain. This requires genotyping of cohorts of mice separated from the main stock at known dates. The
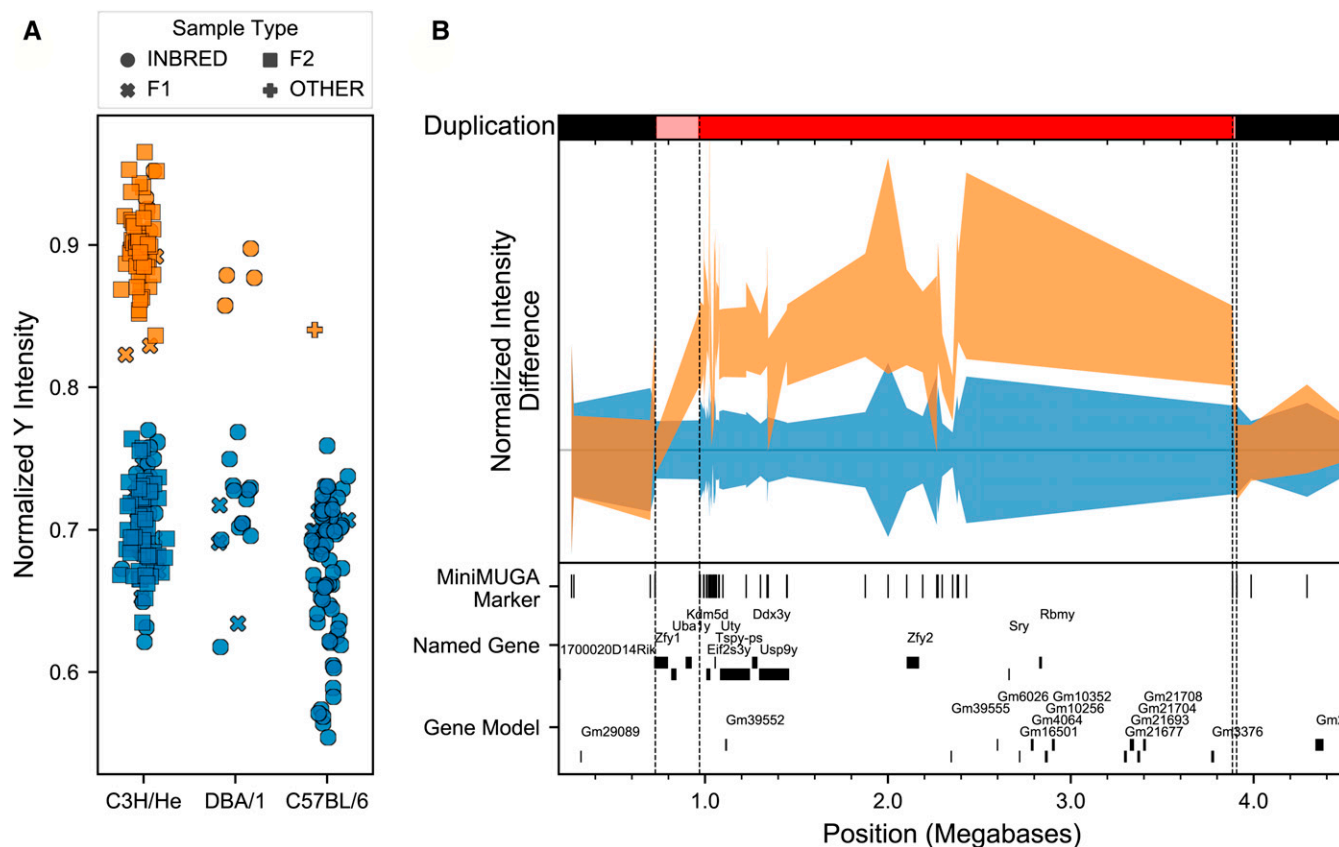
Figure 4 Segmental chromosome *Y* duplications in laboratory strains. (A) Normalized median *Y* chromosome intensity in selected samples with C3H/He, DBA/1, and C57BL/6 *Y* chromosomes. Within the C3H/He group, samples with a C3H/HeJ *Y* chromosome are shown in orange while samples with any other C3H/He *Y* chromosome are shown in blue. For DBA/1, there are multiple technical replicates of a single sample with abnormally high intensity shown in orange. For C57BL/6, there is only one sample with abnormally high intensity. The shape of the point reflects the type of mouse. (B) Range of normalized intensity distributions located at 63 SNPs on the short arm and the beginning of the long arm of *Y* chromosome in the C3H/He samples shown in (A). The range of intensities (mean ± SD) in samples with a C3H/HeJ *Y* chromosome are shown in orange while samples with any other types of C3H/He *Y* chromosomes are shown in blue. At the top of the panel, the potential duplication is shown in red, transition regions with uncertain copy number are shown in pink, and normal copy numbers are shown in black. The bottom of the panel shows the location of the MiniMUGA markers and genes. MUGA, Mouse Universal Genotyping Array.

confidence of the inferences will depend on the size of those cohorts. Those cohorts can be historical samples or extant inbred strains derived at known dates from one or more substrains, such as panels of recombinant inbred lines (RILs), congenics, and consomics.

We have two such panels in our sample set: the BXD and the CC RIL. In the former, we determined whether diagnostic alleles for C57BL/6J and DBA/2J were present in 49 BXD RILs. These RIL were generated in three different epochs: 22 of the genotyped BXD lines belong to epoch I (Taylor *et al.* 1973), four belong to epoch II (Taylor *et al.* 1999), and 23 belong to epoch III (Peirce *et al.* 2004). In the CC population we determined whether diagnostic SNPs for C57BL/6J, A/J, 129S1/SvImJ, and NOD/ShiLtJ were present in 72 CC RILs. CC strains were generated in two waves and at three independent sites from inbred mice originally obtained from The Jackson Laboratory in 2004 and 2007 (Collaborative Cross Consortium 2012). For each SNP, we determined in which relevant cohort the diagnostic allele was first observed, and if

and when it became fixed. This analysis depended on the number of cohorts relevant for a given substrain and the number of samples per cohort. Note that the analysis for a given substrain may integrate multiple cohorts from different populations as long as the year of origin is known. For the five substrains analyzed here, there is considerable variation in the number of cohorts and samples (Table 4). Note that only diagnostic SNPs included in the preliminary phase were used in this analysis because most CC and BXD samples were only genotyped with that version of the array. Also note that the number of independent samples used to establish the consensus is critical to gauge the strength of support for date of fixation of diagnostic alleles in that cohort (Tables S1, S4, and S5). Finally, in the analyses involving the consensus cohorts, we excluded 33 samples because they represent DNA acquired from The Jackson Laboratory >10 years ago.

C57BL/6J is the only shared parental strain in the BXD and CC panels, it is also the most popular inbred strain for experimental biologists, and it is the basis for the mouse reference
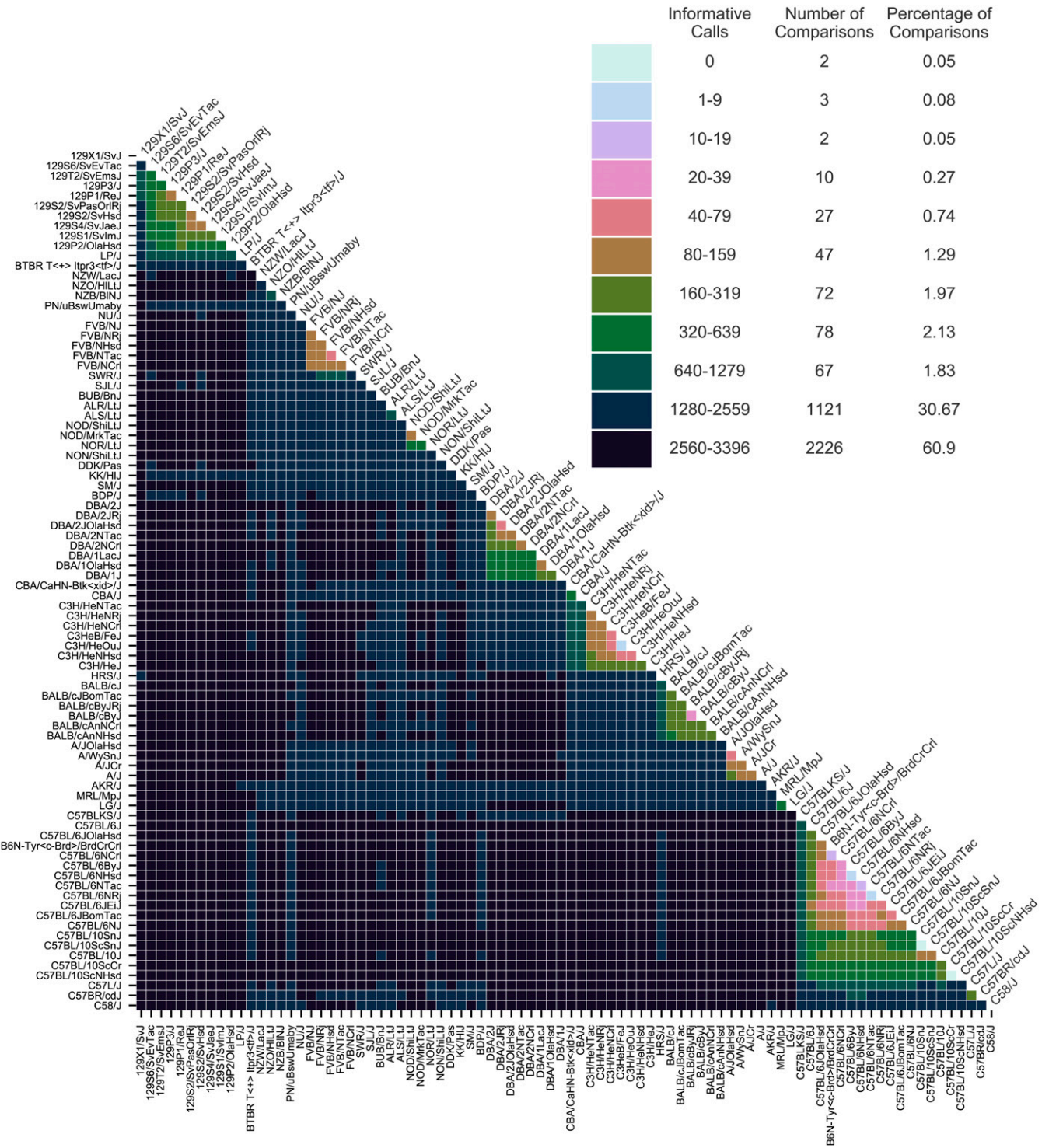
Figure 5 Number of informative SNP calls in pairwise comparisons among classical inbred strains. Strains are ordered by similarity and colors represent the range of number of informative SNPs based on the consensus genotypes. Only homozygous base calls, at tier 1 and 2 markers, on the autosomes, *X*, and pseudoautosomal region are included.

genome. Therefore, we selected C57BL/6J as an example for the procedure and utility for dating diagnostic alleles. The 156 C57BL/6J diagnostic markers were classified based on the earliest observation and apparent fixation in Table 5. Notably, 141 SNPs were distributed in 8 of the 15 possible birth/

fixation pairwise configurations (Table 5). The remaining 15 SNPs were segregating in the most recent cohort, with one half of them segregating since 2004. These SNPs probably represent variants present in the original pair used in the genetic integrity program at The Jackson Laboratory (Sarsani
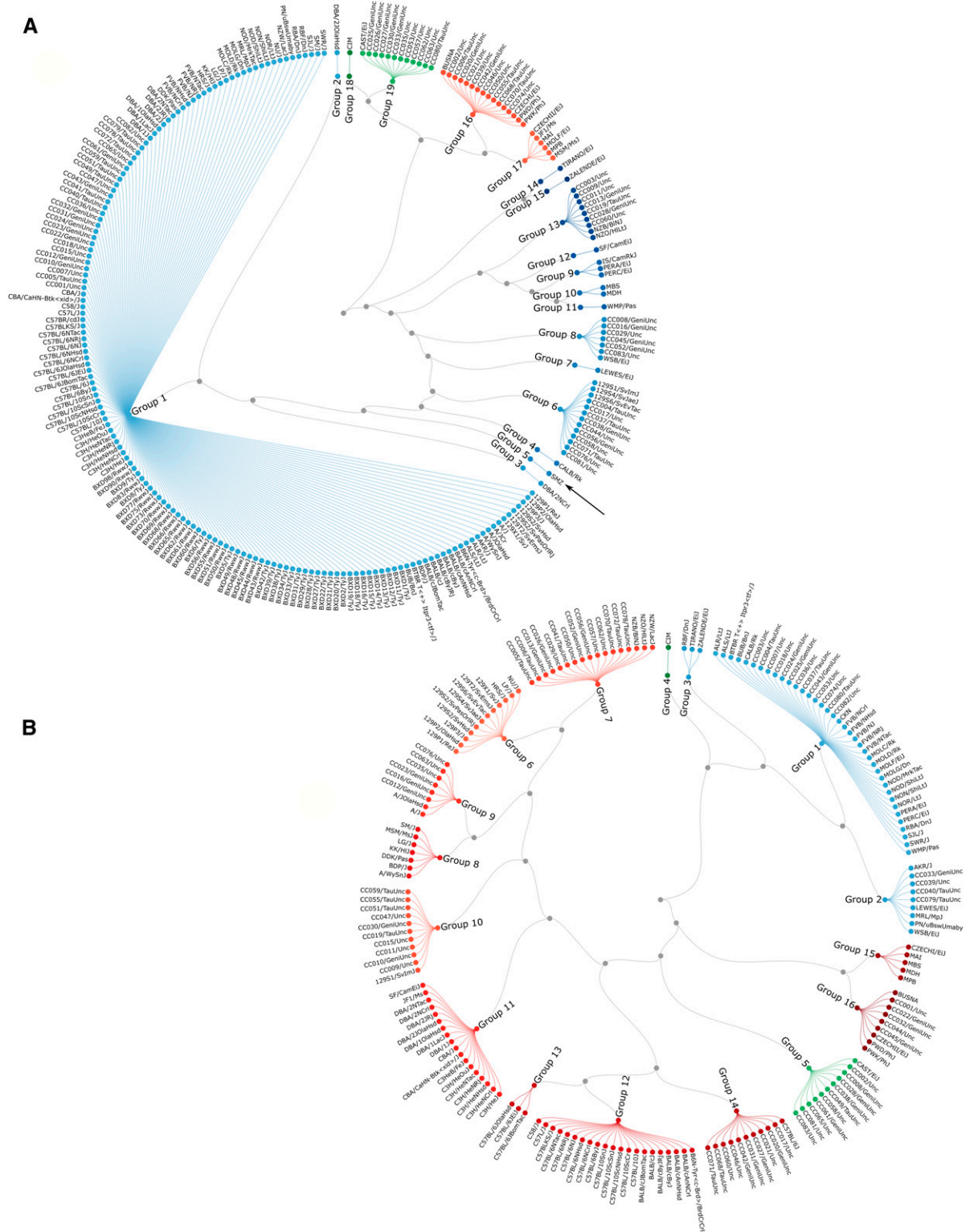
Figure 6 Haplotype diversities. Haplotype diversities of the mitochondria (A) and chromosome *Y* (B). The trees are built based on the variation present in MiniMUGA and may not represent the real phylogenetic relationships. Colors denote the subspecies-specific origin of the haplotype in question: shades of blue represent *M. m. domesticus* haplotypes; shades of red represent *M. m. musculus* haplotypes; and shades of green represent *M. m. castaneus* haplotypes. The arrow in panel (A) identifies a *M. spretus* strain with a *M. m. domesticus* mitochondria haplotype. MUGA, Mouse Universal Genotyping Array.

Table 4 Dating the origin and fixation of diagnostic SNPs in five mouse inbred strains

Substrain	Cohort	Year	Number of samples	Range of alleles sampled	Diagnostic allele		
					Absent	Segregating	Fixed
C57BL/6J	BXD E1	1971	22	11	156	0	0
	BXD E2	1996	4	0–4	84	72	0
	BXD E3	2001–2009	24 (23)	11.5	50	31	75
	CC	2004–2007	483 (72)	4–18	8	30	118
	Consensus[a]	2010–2016	15 (1)	15	0	20	136
DBA/2J	BXD E1	1971	22	11	105	7	0
	BXD E2	1996	4	0–4	37	62	13
	BXD E3	2001–2009	24 (23)	11.5	24	75	13
	Consensus[a]	2010–2016	3 (1)	3	0	0	112
A/J	CC	2004–2007	483 (72)	2–22	2	11	47
	Consensus[a]	2010–2016	10 (1)	10	0	5	55
129S1/SvImJ	CC	2004–2007	483 (72)	3–42	1	6	81
	Consensus[a]	2010–2016	10 (1)	10	0	4	84
NOD/ShiLtJ	CC	2004–2007	483 (72)	4–43	1	2	34
	Consensus[a]	2010–2016	8 (1)	8	0	1	36

This table lists the name of the substrain, the cohorts used for dating the diagnostic SNPs, the approximate year(s) when these cohorts were derived from the main stock, the number of samples genotyped, and the range of alleles sampled. When the number of samples does not match the number of strains, the number of strains is shown in parentheses. Diagnostic alleles are classified as absent, segregating, and fixed for each substrain and cohort, and the table provides the total number in each category.
[a] In this analysis, we excluded samples purchased from The Jackson Laboratory (the sample names include the suffix jaxDNA) over a decade ago in the consensus cohorts. Details are provided in the text. CC, Collaborative Cross; BXD, recombinant inbred BXD panel

et al. 2019). The dates of origin and fixation for C57BL/6J, A/J, 129S1/SvImJ, NOD/ShiLtJ, and DBA/2J diagnostic SNPs are provided in Table S2.

The birth and fixation of diagnostic alleles can be used to determine the origin and breeding history of a given sample of the appropriate background, and thus estimate the expected level of drift (see *Discussion*). The presence of segregating variants for C57BL/6J, A/J, 129S1/SvImJ, and NOD/ShiLtJ at the initiation of the CC project will result in CC strains that share identical haplotypes but may be functionally different due to one of those variants, as has been observed for gene deletion in the BXD panel (Anderson *et al.* 2002; Mulligan *et al.* 2012).

To test whether it is possible to use the diagnostic SNPs with known dates of origin and fixation (Table S2) to determine the breeding history of a given sample or stock, we selected the 156 C57BL/6J diagnostic SNPs as a test case. The key step in this analysis is to identify all SNPs in the sample that have the ancestral allele at fixed diagnostic SNPs. These SNPs identify genomic regions in that sample that have not been

refreshed since fixation of the SNPs. Conversely, SNPs with the derived (diagnostic) allele identify regions in the sample that have been in contact with the main stock since the date of origin of that derived allele. Figure 7 shows the result of this analysis in three samples with different patterns. Figure 7A shows a knockout mouse from a line created prior to epoch III of the BXD panel and bred independently from the C57BL/6J stock since at least 2004. The former conclusion is based on the fact that we detect the ancestral allele at 21 SNPs that were fixed in the C57BL/6J stock prior to epoch III (Table 5). The latter is based on the observation of ancestral alleles at 36 SNPs that were fixed by 2004 (Table 5) and that these markers are distributed across 14 chromosomes. Figure 7B shows a transgenic mouse from a line created prior to the initiation of the CC (2004) and bred independently from the C57BL/6J stock since them. Both conclusions are based on the fact that there are zero ancestral alleles at any of 75 diagnostic SNPs fixed by epoch III (Table 5), the detection of the ancestral allele at 18 SNPs that were fixed prior to the CC (Table 5), and that these markers are distributed across

Table 5 Full dating of diagnostic alleles for the C57BL/6J substrain

		Apparent fixation					Not fixed
		BXD E1	BXD E2	BXD E3	CC	Consensus[a]	
Earliest observation	BXD E1	0	0	0	0	0	0
	BXD E2	NA	0	67	1	4	0
	BXD E3	NA	NA	8	18	8	0
	CC	NA	NA	NA	24	4	14
	Consensus[a]	NA	NA	NA	NA	2	6

The table classifies 156 diagnostic SNPs into one of 20 categories based on the earliest observation (origin) and apparent date of fixation based on whether the diagnostic allele is observed in BXD and CC strains with the C57BL/6J haplotype at each loci. Temporally impossible cells are shown as NA. BXD, Recombinant inbred BXD panel; CC, collaborative cross.
[a] In this analysis, we excluded samples purchased from The Jackson Laboratory (the sample names include the suffix jaxDNA) over a decade ago in the consensus cohorts. Details are provided in the text.
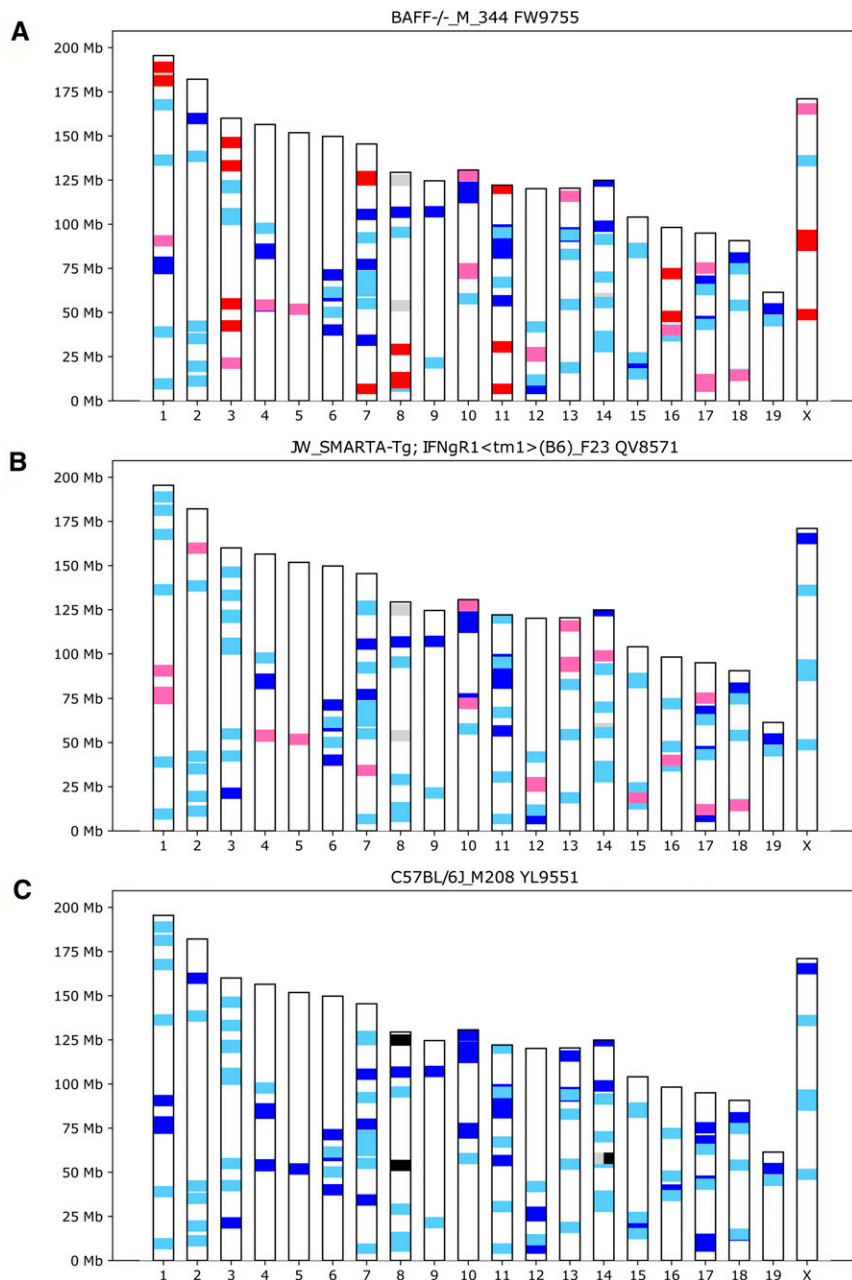
Figure 7 Sample dating and breeding history of mice with C57BL/6J background. Red bars denote the ancestral allele for diagnostic SNPs fixed by E3 in the BXD panel. Pink bars denote ancestral alleles for diagnostic SNPs fixed by the start of the CC. Light blue bars denote diagnostic alleles at diagnostic SNPs fixed by E3. Blue bars denote diagnostic alleles at diagnostic SNPs fixed by the start of CC. Gray bars denote ancestral alleles at post-CC diagnostic SNPs. Black bars denote diagnostic alleles at post-CC diagnostic SNPs. Split bars denote heterozygosity. (A) Inbred *Baff*$^{-/-}$ male in C57BL/6J background. (B) Inbred transgenic and IFNgR1 female in C57BL/6J background. (C) Inbred C57BL/6J male. Diagnostic allele always represent the derived allele, and the nondiagnostic allele is always the ancestral allele. CC, collaborative cross; BXD, Recombinant inbred BXD panel; E, epoch.

13 chromosomes. Finally, Figure 7C shows a wild-type C57BL/6J mouse derived from the JAX colony after 2004. This conclusion is based on the lack of ancestral alleles at any of 124 fixed diagnostic SNPs and the presence of a derived allele at three SNPs that arose after the CC (Table 5). Notably, these conclusions are consistent with the expectations of the contributors of these samples.

Expansion of reduced complexity crosses to a large number of substrains

We define RCC as crosses between substrains derived from a single inbred strain that differed only at mutations that arose after they were isolated and bred independently from a common stock. We tested the ability of MiniMUGA to efficiently cover the genome in 78 different RCC between substrains for which we had consensus genotypes, WGS, and for which live mice were available from commercial vendors (see Table 2). We focused our analysis on this group given that WGS of both substrains is required for rapid identification of causative variant(s) (Kumar *et al.* 2013; Treger *et al.* 2019). We used the distance to the nearest informative marker to estimate how well MiniMUGA covers the genome in a given RCC cross. Figure 8 and Table S9 summarize these data, and demonstrate that for 62 RCC (82%) all of the genome is covered by a linked marker and in 14 RCC (18%) between 95 and 99.5% of the genome is covered by a linked marker.
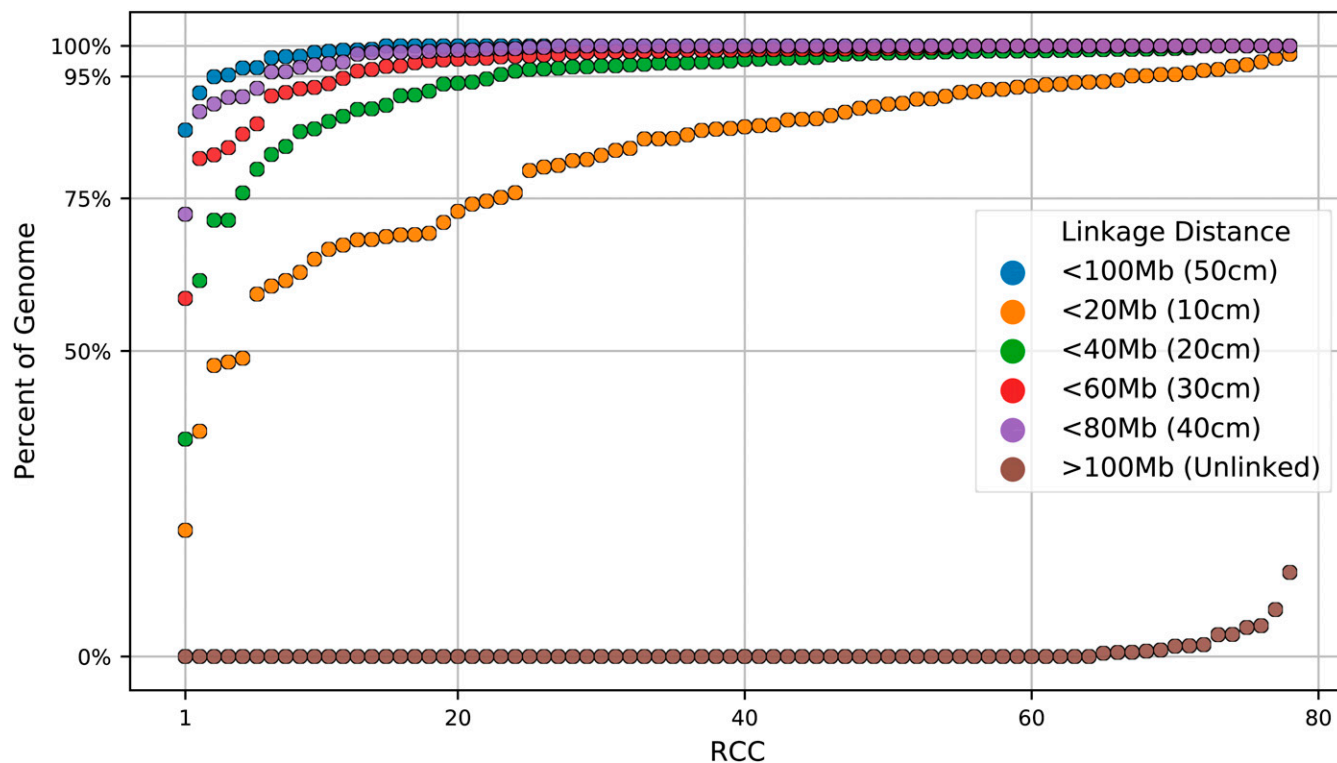
Figure 8 Percent of the genome covered by MiniMUGA in RCC. Each of the 78 RCC is shown as a circle in ascending order. The order is independent for each one of the six analyses. Coverage was based on the linkage distance to the nearest informative marker in given RCC. MUGA, Mouse Universal Genotyping Array; RCC, reduced-complexity crosses.

Only in two RCC (3%) is there a significant fraction of the genome that is not covered by a linked marker. These two crosses are B6N-Tyr < c-Brd>/BrdCrCrl by C57BL/6JOlaHsd and BALB/cByJ by BALB/cByJRj with 8 and 14% of the genome not covered, respectively. An alternative test is the number of RCC for which 95% of the genome is covered by informative markers at 20-cM (56 RCC or 72%) and 40-cM (72 RCC or 92%) intervals. We conclude that MiniMUGA provides a cost-effective tool to extend RCC to substrains from the 129P, 129S, A, BALB/c, C57BL/6, C3H, DBA/1, DBA/2, FVB, and NOD strain groups.

Robust detection of common genetic constructs

Given the broad usage of genetic editing technologies, a key design criterion of MiniMUGA was the ability to detect frequently used genetic constructs. Utilizing our pipeline (see *Materials and Methods*), we positively identified samples containing 17 construct types (Figure 9). Importantly, for eight of these constructs, our sample set also included positive controls. These positive controls showed robust detection of their relevant constructs. We detected further positive samples in our set for these eight constructs, as well as nine additional constructs without positive controls. The latter set of samples belonged to sample classes where constructs were plausible (*e.g.*, not wild-derived or CC samples), and there was high concordance for intensities among the multiple probes of a single construct (Figure S2B intensity correlation).

Across these 17 constructs, we observed that our ability to discriminate between negative and positive samples was strongly correlated with the number of independent probes for that construct (Figure 9 and Figure S2B). As signal intensity is constrained by the dynamic range, our ability to definitively call the presence of low-probe number constructs is more uncertain. This uncertainty is especially relevant where a construct within a sample is genetically divergent from the sequences used to design a given probe/probe set. Given our ability to positively identify construct classes with as few as two probes, it is likely that even for constructs which have divergent sequences from our designed sequences, or are targeting a more distantly related construct type, our pipeline will flag samples. However, users are highly encouraged to consult the probe sequences (Figure S3) when they expect a given sample to contain a construct, but do not see support in the array itself. Conversely, if a construct with many independent probes is determined to be present, that call is more reliable, even if a sample is not expected to contain that construct.

We additionally observed that for some constructs, there was between-sample variation in the overall intensity of the signal associated with a given construct [see internal ribosome entry site (IRES), Figure 9]. For IRES, the between-sample variation was likely to be due a higher copy number of the construct in five individual samples because of consistent higher intensity across all probes (Figure S2A). Copy-number
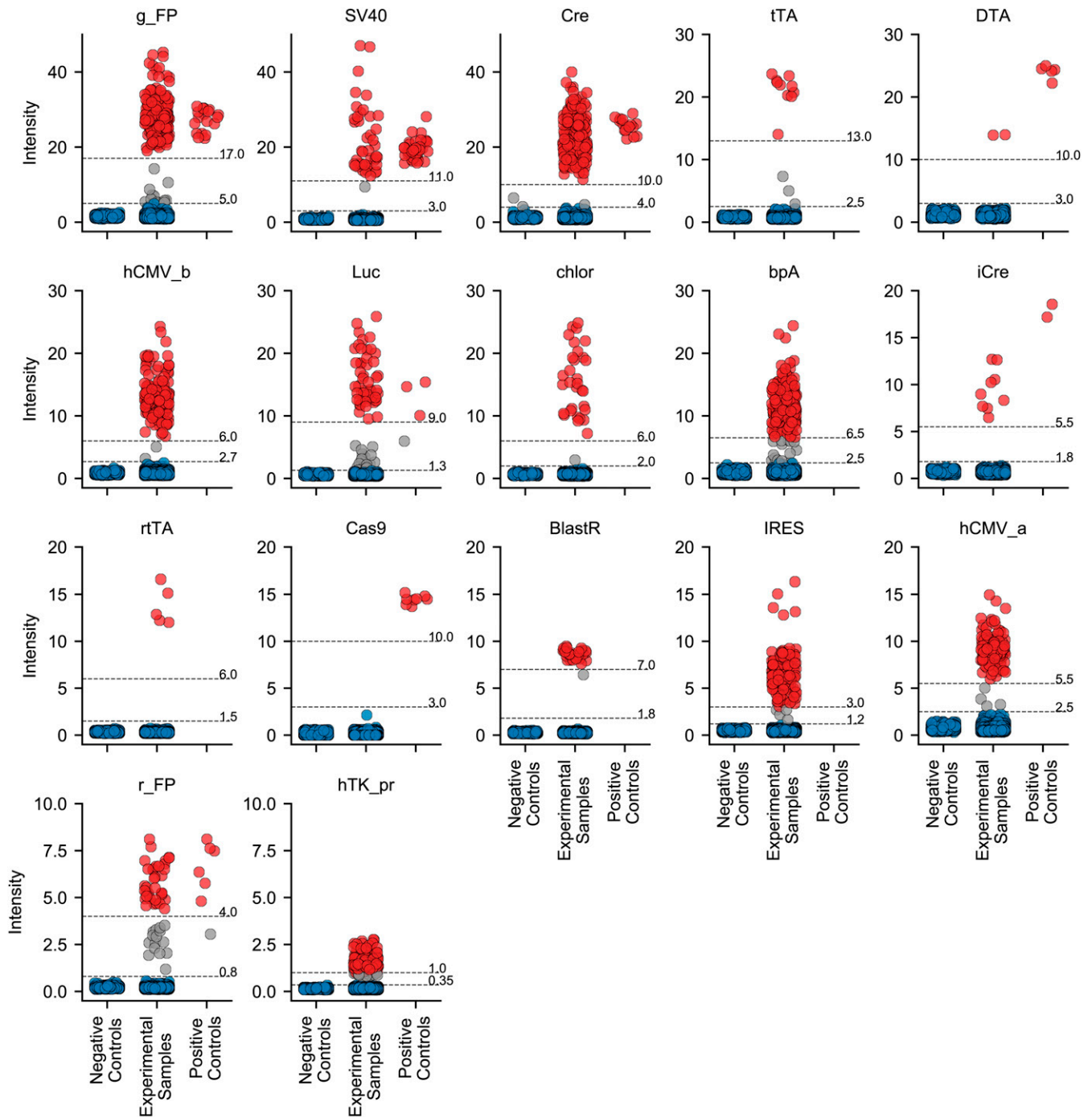
Figure 9 Detection of genetic constructs validated in MiniMUGA. For each construct, samples are shown as dots and classified as negative controls (left), experimental (center), and positive controls (right). The dot color denotes whether the sample is determined to be negative (blue), positive (red), or questionable (gray) for the respective construct. For each construct, the gray horizontal lines represent data-driven *ad hoc* thresholds discriminating between presence and absence. Note for each construct, the y-axis scale is different. MUGA, Mouse Universal Genotyping Array; g_FP; 'greenish' fluorescent protein; SV40; SV40 large T antigen; Cre, Cre recombinase; tTA, tetracycline repressor protein;; DTA, Diptheria toxin; hCMV_b, Human CMV enhancer version b; Luc, Luciferase and firefly luciferase; chloR, Chloramphenicol acetyltransferase; bpA, Bovine growth hormone poly A signal sequence; iCre, iCre recombinanse; rtTA, Reverse improved tetracycline-controlled transactivator; cas9, CRISPR associated protein 9; BlastR, Blasticidin resistance; IRES, Internal Ribosome Entry Site; hCMV_a, hCMV enhancer version a; r_FP, 'reddish' fluorescent protein; hTK_pr, Herpesvirus TK promoter.

variation in transgene insertions is a common phenomenon (see "Development" documentation on JAX Stock 034860, McCray *et al.* 2007), and such copy-number variation segregating within a given colony/line can lead to noise and a lack of reproducibility in given experiments. Alternatively, between-sample variation might be explained by

MiniMUGA Background Analysis v0008

Sample ID	MMRRC_UNC_F38673
Neogen ID	US7600
Summary	The genotype of this sample is of **excellent** quality. It is **female** and **close to inbred**, and likely a mix of **multiple C57BL/6 substrains** and **(129S1/SvImJ and/or 129S2/SvHsd and/or 129S2/SvPasOrlRj and/or 129S4/SvJaeJ and/or 129S6/SvEvTac)**. Clustering of unexplained markers is evidence of an additional background strain. Diagnostic SNPs indicate the presence of the background strain groups **C57BL/6** and the substrains **C57BL/6J**. The sample contains the following genetic constructs: **Luciferase**
Genotyping Quality	**Excellent (18 N calls)** All reported results are dependent on genotyping quality.
Chromosomal Sex	XX
Inbreeding Estimate	**Close to Inbred (200 H calls at autosomal, X, and PAR chromosome markers)**

Inbreeding and Genotyping Quality (Plot)

Constructs Detected

BlastR	bpA	Cas9	chloR	Cre	DTA	g_FP	hCMV_a	hCMV_b	hTK_pr	iCre	IRES	Luc	r_FP	rtTA	SV40	tTA
-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-

Primary Background (Autosomes, X Chromosome)

Strain	Total	Consistent	Inconsistent	Heterozygous	Excluded
multiple C57BL/6 substrains	9721	9087 (97.9%)	50 (0.5%)	148 (1.6%)	436

Secondary Background (Autosomes, X Chromosome)

Strain	Total	Explained	Unexplained	Excluded
129S1/SvImJ and/or 129S2/SvHsd and/or 129S2/SvPasOrlRj and/or 129S4/SvJaeJ and/or 129S6/SvEvTac	198 193 Clustered	182 (2.0%) 181 Clustered	16 (0.2%) 7 Clustered	0 (0.0%)

Background Ideogram

Backgrounds Detected (Diagnostic Alleles)

	Diagnostic Alleles Observed			
Substrain	Homozygous	Heterozygous	Potential	% Observed
C57BL/6J	77	45	156	78.2%
Strain Group	Homozygous	Heterozygous	Potential	% Observed
C57BL/6 (B6N-Tyr/BrdCrCrl, C57BL/6J, C57BL/6JBomTac, C57BL/6JEiJ, C57BL/6JOlaHsd, C57BL/6NCrl, C57BL/6NHsd, C57BL/6NJ, C57BL/6NRj, C57BL/6NTac)	6	1	21	33.3%

Figure 10 Background Analysis Report for the sample named MMRRC_UNC_F38673, from the line named B6.Cg-*Cdkn2a*$^{tm3.1Nesh}$ *Tyr*$^{c−2J}$ *Hr*^{hr}/Mmnc. The genotype of this sample is of excellent quality. It is a close to inbred female that is a congenic with C57BL/6J as a primary background, and with multiple regions of a 129S secondary background. This sample is positive for the luciferase and firefly luciferase construct, and negative for 16 other constructs. g_FP; 'greenish' fluorescent protein; SV40; SV40 large T antigen; Cre, Cre recombinase; tTA, tetracycline repressor protein;; DTA, Diptheria toxin; hCMV_b, Human CMV enhancer version b; Luc, Luciferase and firefly luciferase; chloR, Chloramphenicol acetyltransferase; bpA, Bovine growth hormone poly A signal sequence; iCre, iCre recombinanse; rtTA, Reverse improved tetracycline-controlled transactivator; cas9, CRISPR associated protein 9; BlastR, Blasticidin resistance; IRES, Internal Ribosome Entry Site; hCMV_a, hCMV enhancer version a; r_FP, 'reddish' fluorescent protein; hTK_pr, Herpesvirus TK promoter; PAR. pseudoautosomal region

between-probe variation for a construct as discussed above. Individuals are encouraged to examine the summed intensity levels for positive constructs for their strains/samples, to confirm that within a relevant sample group, these levels are roughly equal.

An easy-to-interpret report summarizes the genetic QC for every sample

The MiniMUGA Background Analysis Report (Figure 10) aims to provide users with essential sample information derived from the genotyping array for every sample genotyped. The report is designed to provide overall sample QC, as well as genetic background information for classical inbred mouse strains, and congenic and transgenic mice. For samples outside of this scope, the report may be incomplete and/or provide misleading conclusions. Details of the thresholds and algorithms for each section of the report are provided in the *Materials and Methods* section.

In addition to chromosomal sex and the presence of constructs, the report provides a quantitative and qualitative score for genotyping quality. Based on the number of N calls per sample of our sample set, we classified samples in one of four categories: samples with excellent quality (0–91 N calls, represents 96.8% of samples), samples with good quality (between 92 and 234 N calls, 2% of samples), samples with questionable quality (between 235 and 446 N calls, 0.9% of samples), and samples with poor quality (>447 N calls, 0.3% of samples). Only tier 1 and 2 markers were used in this analysis (see *Materials and Methods*).

Regarding inbreeding status, the report assigns every sample to one of three categories: Inbred (<61 H calls), close to inbred (between 61 and 280 H calls), and outbred (>280 H calls). These thresholds are based on the number of H calls observed in the autosomes of 172 samples of classical inbred strains and predicted heterozygosity in 3655 *in silico* F_1 hybrid mice (Figure S7).

For genetic background detection, the report provides two complementary analyses. The first infers the primary and secondary backgrounds of samples that pass genotype quality and inbreeding thresholds based on the totality of their genotypes (excluding the *Y* chromosome, mitochondria markers, and construct probes). The second returns the genetic backgrounds detected in a sample based on the presence of the diagnostic allele at diagnostic SNPs (see section on *Diagnostic SNPs as a tool for genetic QC and strain dating*).

For the primary background analysis, the sample's genotype is compared to a set of 120 classical and wild-derived inbred reference strains (Table S3) to identify the strain that best explains the sample genotypes. If multiple substrains from the same strain group have been detected via diagnostic alleles, or if there is an overrepresentation of a particular diagnostic strain in the unexplained markers, the algorithm generates a composite strain consensus that incorporates all substrains in that strain group and uses it in the primary background analysis. The strain or combination of substrains that best matches the sample is called the primary

background for the sample. The report provides the number of homozygous calls that are consistent or inconsistent with the primary background, as well as the number of heterozygous calls in the sample. The primary background is always returned for samples in which the primary background explains at least 99.8% of the sample genotype calls.

Once the primary background is identified, the algorithm tests whether ≥75% of the markers inconsistent with the primary strain background and heterozygous markers are spatially clustered. If they are not (<75% of markers spatially clustered) the algorithm will not try to identify a secondary background. If ≥75% of the unexplained markers are clustered, all reference strains that equally explain the unexplained calls are identified as potential secondary background(s). If the combination of primary and secondary backgrounds explains ≥99.8% of the calls, the primary and secondary backgrounds are reported. If this combination explains <99.8%, then no genetic background is returned.

For samples where a primary and secondary background is reported, the algorithm determines whether the remaining unexplained markers are spatially clustered. If they are, the summary states that clustering of unexplained markers may indicate the presence of an additional genetic background. The limitations of this greedy approach to identification of the primary and secondary backgrounds are addressed in the *Discussion* section.

Note that this report is generated programmatically using an available set of reference inbred strains (Table S3). If the reported results are inconsistent with expectations, users need to consider further analyses before reaching a final conclusion. All estimates and claims in the report are heavily dependent on the quality of the sample and genotyping results. Less than excellent genotyping quality may increase the likelihood of an incorrect background determination. Genotyping noise can lead to incorrect reporting and may be particularly misleading in samples from standard commercial inbred strains. Fully inbred strains routinely have a small percentage of spurious H calls. These do not represent true heterozygosity (see consensus of inbred strains).

Discussion

MiniMUGA as a tool for QC

Among the many new capabilities of the MiniMUGA array compared with its predecessors is the Background Analysis Report provided with each genotyped sample. Although expert users can, and undoubtedly will, refine existing and develop new analyses pipelines, all users benefit from a common baseline developed after the analyses of thousands of samples. The size, annotation, and variety of our sample set provides a firm foundation for our conclusions.

We urge users to pay particular attention to genotype quality, reported heterozygosity, and unexpected conclusions (*i.e.*, sex, backgrounds, and constructs detected). Genotype

quality depends on the sample quality, quantity, and purity, and on the actual genotyping process. Poor genotype quality can also be the byproduct of off-target variants in the probes used for genotyping and, thus, wild mouse samples and mice from related taxa are expected to have lower apparent quality (Didion *et al.* 2012). Samples with poor quality will not be run through the full report pipeline. Samples with questionable quality may lead to incorrect conclusions. For samples of any quality the total number of N calls should be carefully considered if unexpected results are reported. It is also important to consider a sample's probe intensity distribution value as represented by pd_stat when evaluating the credibility of its reported chromosomal sex. The reported chromosomal sex of samples with high pd_stat (>3230) should be questioned (see Figure S5).

Reported heterozygosity is sensitive to genotyping quality. A lower-quality sample will typically include more spurious heterozygous calls than an excellent-quality sample of the same strain. This leads to an incorrect estimate of the level of inbreeding in a given sample and can be particularly misleading in a fully inbred mouse of a single background. The thresholds used to classify samples as inbred, close to inbred, and outbred are somewhat arbitrary and reflect the biases in SNP selection (overrepresentation of diagnostic SNPs for selected substrains) and the highly variable range of diversity observed in F1 mice. We used the observed number of H calls in known inbred samples and the predicted number of H calls among a large and varied set of potential F1 hybrids to set our thresholds, but users should consider the level of heterozygosity expected in a specific experiment (Figure S7). For example, mice generated in RCC between related substrains may have a very small number of H calls and thus will be misclassified as more inbred than they really are. The report combines sample quality and heterozygosity in a single figure for quick visual inspection (Figure 10). Note that the *x*- and *y*-axes are compressed in the high-value range to ensure that all samples, even those with very poor quality and/or high heterozygosity, are shown. The precise location of a sample in the plot should help customers contextualize their sample's quality and inbreeding when evaluating their results.

For users genotyping a large number of samples in a given batch (for example, several 96-well-plates), we found it useful to include a plate-specific control at an unambiguous location (we used the B3 well). Ideally, these controls should have known genotypes, excellent quality, and be easy to distinguish from all other samples in the batch. Plating errors or unaccounted transpositions occurring during the genotyping process are rare but problematic. Adding one sample per plate is a reasonable price to pay to quickly identify these issues.

We anticipate that most users will use the Background Analysis Report to determine the genetic background(s) present in a sample as well as their respective contributions. The identification of the correct primary and secondary background is completely dependent on the preexisting set of reference strains (Table S3). If a genotyped sample is derived from a strain that is not part of this reference set, the reported results may be misleading or completely incorrect. Users should consult the list of reference backgrounds (Table S3). We expect the number of reference backgrounds to increase over time, reducing the frequency and impact of this problem. However, the current background detection pipeline is not appropriate for RILs such as the BXD and CC populations. By their very nature, these RILs have mosaic genomes derived from two or more inbred strains included in our panel, and thus the background analysis will detect more than two inbred backgrounds for CC strains, or declare one of the parental strains as primary or secondary background for the BXD strains. Users interested in confirming or determining the identity of RILs can use our consensus genotypes to do so.

An important caveat of the current primary and secondary background analysis is that the approach is greedy, and all variants except those with H and N calls in the consensus are considered. Because only a fraction of the SNPs are informative between a given pair of strains (always less than one half, see Figure 5 and Figure S7), the algorithm always overestimates the contribution of the primary background and underestimates the contribution of the secondary background (Figure 10). In congenic strains, the true contribution of the strain identified as the secondary background is approximately 2.6 times higher than shown in the report (Figure S9). This appears to be true independent of the strains identified as primary and secondary backgrounds and their proportions. If the exact contribution of either background is critical for the research question, the user should reanalyze the data using only SNPs that discriminate between the two backgrounds.

A second caveat is that the current pipeline does not include the mitochondria and *Y* chromosomes. This shortcoming will be addressed in a future update of the Background Analysis Report.

A final caveat is that in most cases where more than two inbred strains are needed to explain the genotypes of a sample, the report does not identify any of them. In our experience, when three of more backgrounds are present, a greedy search is not effective and often leads to incorrect results. Therefore, if the user has prior knowledge of at least some of the backgrounds involved, they should conduct an iterative hierarchical search that will typically yield the correct solution, but care needs to be taken at each step.

The private variants that underlie the RCC concept are the diagnostic variants used in background determination and sample dating. Diagnostic SNPs have little information content but high specificity. The presence of diagnostic alleles in a sample is strong evidence that that specific substrain (or a closely related substrain absent from our set) contributed to the genetic background of that sample. However, because only a small fraction of diagnostic SNPs have been observed in all three genotypes across multiple samples, their performance is not well established, in particular for heterozygous calls. To avoid errors, we required diagnostic alleles at three different SNPs in a given sample before a genetic background was declared in the Background Analysis Report. All diagnostic

SNPs began their history as partially diagnostic (segregating in an inbred strain or substrain population).

The examples for dating stocks as shown in the *Results* section were fairly simple, but more complex and more interesting patterns are plentiful in our data set. For example, four samples from a congenic inbred stock showed evidence of both an old stock and new refreshing of the genome in recent years (Figure S10). Specifically, the presence of ancestral alleles at many diagnostic SNPs fixed prior to epoch III and the start of the CC speaks of a mouse line generated and bred independently for many years. On the other hand, heterozygosity at some of these markers as well as the presence of diagnostic alleles that are still segregating (Table 5) indicates that this line was refreshed by backcrossing to C57BL/6J in recent years. Both conclusions are consistent with our expectations as this stock was imported by Mark Heise at UNC in 2014 and backcrossed once or a few times to JAX mice before being maintained by brother–sister mating. In addition to improving the genetic QC, we believe that this type of analysis may provide researchers with critical information to guide both experimental design and data analysis. Also important is the ability to estimate the amount of drift that will occur and thus the amount of genetic variation present in that line but absent in the main stock. We expect that widespread use of MiniMUGA and the continued and rapid annotation of diagnostic SNPs (Table 4), not only for C57BL/6J but for all inbred substrains, offers an opportunity to significantly improve the rigor and reproducibility of mouse research.

Mouse cell lines can be subject to the same genetic QC as mice. We have shown the ability to detect sex chromosome aneuploidy in cell lines (Figure 1). Diagnostic SNPs can be used to date cell lines in similar fashion to live mice with the added simplicity that cell lines are less susceptible to drift. Finally, the Background Analysis Report pipeline can be used to effectively identify the origin of cell lines. Examples are provided in Figure S8. The importance of genetic QC in cell lines will grow in the future given the increased emphasis on cell-based research. We have previously reported that the number of N calls on other genotyping platforms is higher for cell lines than for biopsies (Didion *et al.* 2014). The evidence of such phenomena in our data set is inconclusive.

Genetic constructs have been a staple of genome editing technologies since the 1980s. In addition to desired genetic modifications, constructs will often include a variety of other necessary features (*e.g.*, selection markers and constitutive promoters). The array can be used to validate the presence of constructs expected to be present and/or to identify the presence of unexpected constructs.

Our construct probe design was focused on targeting conserved features of various genetic engineering and/or *in vitro* constructs commonly used in mammalian genetics. We can split these conserved probe sets into two main classes: those for which we were able to detect positive samples in our data set, and those for which we were not able to detect any consistently positive samples. Given

the between-probe variation in a given sample, interested users can examine the individual probe intensities to refine the analysis (*e.g.*, the cyan, green, and yellow fluorescent protein probe sets).

Finally, we designed probes for 14 constructs (123 probes) for which we were unable to call presence or absence in our pipeline. This may be due to lack of positive samples in our data set, not enough probes with positive signal for a given construct, or probes that failed. If a user knows that a construct is present in their data set, they are encouraged to recreate our pipeline for calling presence or absence of the relevant construct.

MiniMUGA as a tool for discovery

MiniMUGA was designed to support the research mission of geneticists, but the range of applications will depend on the ingenuity of its users. In the *Results* sections, we explored three areas in which MiniMUGA has the potential to enhance existing resources and tools.

The first of these areas was sex chromosome biology. MiniMUGA is able to robustly determine four sex chromosome configurations (Figure 1) and thus facilitate estimation of the incidence and prevalence of sex chromosome aneuploidy in the mouse. The variation of aneuploidy rates depending on the sire background provides a promising avenue to study the genetics of sex chromosome missegregation. In addition, identification of aneuploid mice can become routine in experimental cohorts and crosses. This is also important in colony management, as *X*O and *XXY* mice are likely to be infertile or have reduced fertility (Heard and Turner 2011).

This type of analysis can also identify sex chromosome mosaicism (Johnson *et al.* 2010; Fragouli *et al.* 2011; McCoy 2017) and large structural variants involving the sex chromosomes. In the *Results* section, we have shown that mosaics are outliers with respect to the four defined clusters observed in the intensity-based chromosome sex determination plot (Figure 1). Specifically, they have abnormal *Y* chromosome intensities. These mosaics may also have an abnormally high ratio of N calls in the *X* chromosome compared with the autosomes and chromosome *X* marker intensity distributions biased toward one parent (Figure 3). This last analysis is only possible in the presence of heterozygosity on the *X* chromosome.

Further evidence of the value of the MiniMUGA array for the characterization of the sex chromosomes is the identification of a 6-Mb *de novo* duplication of the distal chromosome *X* (Figure S11) in a single F2 male. The size of this duplication is not large enough to affect chromosomal sex determination and its discovery was due to the presence of 10 heterozygous calls clustered on distal *X*. These heterozygous calls occur at informative markers between the two parental CC strains of the F2 cross and are embedded in a region of 26 consecutive markers with higher-than-expected intensity (Figure S11). Interestingly, the parental CC strains (CC029/Unc and CC030/GeniUnc) are the same for which a 10× increase in sex chromosome aneuploidy is observed. We concluded that

this F2 male had a well-defined duplication of the distal X chromosome. These vignettes provide a potential blueprint that can be extended to other chromosomes and structural variants. It also highlights the importance of having a large set of well-defined genotyped controls, against which to compare a given sample.

A second area of potential research is the expansion of the RCC paradigm beyond the narrow confines of C57BL/6 substrains (Kumar *et al.* 2013; Babbs *et al.* 2019; Treger *et al.* 2019). A successful RCC requires complete knowledge of the sequence variants shared and private to the set of substrains that will be used in the mapping experiments. These private variants are needed to infer causation but also in the initial step of genetic mapping. We acknowledge that the development of MiniMUGA was made possible by the efforts of the community to sequence an increasing number of inbred strains. The expansion of RCC to 129S, A, BALB/c, C57BL/6, C3H, DBA/1, DBA/2, FVB, and NOD substrains should increase the total number of accessible private mutations by at least one order of magnitude as compared with RCC involving C57BL/6N and C57BL/6J. Therefore, we should expect a similar increase in the number of causative genetic variants. We note that even as substrains continue to accumulate private variants in an unpredictable manner, MiniMUGA will retain its value for genetic mapping, but future WGS will be required to identify those variants.

Genotyping arrays are a powerful, standardized platform with which to characterize the genomic composition of sets of samples. Here, we have described a new mouse genotyping array, MiniMUGA. We have illustrated how the design and performance of MiniMUGA provides a more robust platform for genetic QC (at the sex-chromosome, mouse substrain identity, and genetic construct levels) relative to our previously designed arrays. We have also illustrated examples of how this array can be used for new genetic discovery, including sex chromosome abnormalities, genomic duplications, and also in the expansion of genetic mapping approaches. This array and these associated data highlight the utility of genetic QC for more robust and reproducible science in the mouse, and they are already being widely used by the research community (Smith *et al.* 2019; Gu *et al.* 2020; Yu *et al.* 2020).

Acknowledgments

Literature Cited

Adams, D. J., A. G. Doran, J. Lilue, and T. M. Keane, 2015   The Mouse Genomes Project: a repository of inbred laboratory mouse strain genomes. Mamm. Genome Off. J. Int. Mamm. Genome Soc. 26: 403–412. https://doi.org/10.1007/s00335-015-9579-6

Akeson, E. C., L. R. Donahue, W. G. Beamer, K. L. Shultz, C. Ackert-Bicknell *et al.*, 2006   Chromosomal inversion discovered in C3H/HeJ mice. Genomics 87: 311–313. https://doi.org/10.1016/j.ygeno.2005.09.022

Anderson, M. G., R. S. Smith, N. L. Hawes, A. Zabeleta, B. Chang *et al.*, 2002   Mutations in genes encoding melanosomal proteins cause pigmentary glaucoma in DBA/2J mice. Nat. Genet. 30: 81–85. https://doi.org/10.1038/ng794

Arends, D., S. Heise, S. Kärst, J. Trost, and G. A. Brockmann, 2016   Fine mapping a major obesity locus (jObes1) using a Berlin Fat Mouse × B6N advanced intercross population. Int. J. Obes. (Lond) 40: 1784–1788. https://doi.org/10.1038/ijo.2016.150

Babbs, R. K., J. A. Beierle, Q. T. Ruan, J. C. Kelliher, M. M. Chen *et al.*, 2019   Cyfip1 haploinsufficiency increases compulsive-like behavior and modulates palatable food intake in mice: dependence on Cyfip2 genetic background, parent-of origin, and sex. G3 (Bethesda) 9: 3009–3022. https://doi.org/10.1534/g3.119.400470

Boratyn, G. M., C. Camacho, P. S. Cooper, G. Coulouris, A. Fong *et al.*, 2013   BLAST: a more efficient report with usability improvements. Nucleic Acids Res. 41: W29–W33. https://doi.org/10.1093/nar/gkt282

Bryant, C. D., D. J. Smith, K. M. Kantak, T. S. Nowak, Jr., R. W. Williams *et al.*, 2020   Facilitating complex trait analysis via reduced complexity crosses. Trends Genet. 36: 549–562. https://doi.org/10.1016/j.tig.2020.05.003

Carbonetto, P., R. Cheng, J. P. Gyekis, C. C. Parker, D. A. Blizard *et al.*, 2014   Discovery and refinement of muscle weight QTLs in B6 × D2 advanced intercross mice. Physiol. Genomics 46: 571–582. https://doi.org/10.1152/physiolgenomics.00055.2014

Cheng, H.-H., C.-Y. Ou, C.-C. Tsai, S.-D. Chang, P.-Y. Hsiao *et al.*, 2014   Chromosome distribution of early miscarriages with present or absent embryos: female predominance. J. Assist. Reprod. Genet. 31: 1059–1064. https://doi.org/10.1007/s10815-014-0261-9

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 (Bethesda) 6: 3893–3902. https://doi.org/10.1534/g3.116.035527

Collaborative Cross Consortium, 2012 The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics 190: 389–401. https://doi.org/10.1534/genetics.111.132639

Dallas, J. F., 1992 Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. Mamm. Genome 3: 452–456. https://doi.org/10.1007/BF00356155

Didion, J. P., and F. Pardo-Manuel de Villena, 2013 Deconstructing Mus gemischus: advances in understanding ancestry, structure, and variation in the genome of the laboratory mouse. Mamm. Genome 24: 1–20. https://doi.org/10.1007/s00335-012-9441-z

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. https://doi.org/10.1186/1471-2164-13-34

Didion, J. P., R. J. Buus, Z. Naghashfar, D. W. Threadgill, H. C. Morse et al., 2014 SNP array profiling of mouse cell lines identifies their strains of origin and reveals cross-contamination and widespread aneuploidy. BMC Genomics 15: 847. https://doi.org/10.1186/1471-2164-15-847

Didion, J. P., A. P. Morgan, L. Yadgary, T. A. Bell, R. C. McMullan et al., 2016 R2d2 drives selfish sweeps in the house mouse. Mol. Biol. Evol. 33: 1381–1395. https://doi.org/10.1093/molbev/msw036

Doran, A. G., K. Wong, J. Flint, D. J. Adams, K. W. Hunter et al., 2016 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. https://doi.org/10.1186/s13059-016-1024-y

Egan, C. M., S. Sridhar, M. Wigler, and I. M. Hall, 2007 Recurrent DNA copy number variation in the laboratory mouse. Nat. Genet. 39: 1384–1389. https://doi.org/10.1038/ng.2007.19

Fragouli, E., S. Alfarawati, D. D. Daphnis, N.-N. Goodall, A. Mania et al., 2011 Cytogenetic analysis of human blastocysts with the use of FISH, CGH and aCGH: scientific data and technical evaluation. Hum. Reprod. 26: 480–490. https://doi.org/10.1093/humrep/deq344

Gu, B., J. R. Shorter, L. H. Williams, T. A. Bell, P. Hock et al., 2020 Collaborative Cross mice reveal extreme epilepsy phenotypes and genetic loci for seizure susceptibility. Epilepsia DOI: 10.1111/epi.16617. https://doi.org/10.1111/epi.16617

Heard, E., and J. Turner, 2011 Function of the sex chromosomes in mammalian fertility. Cold Spring Harb. Perspect. Biol. 3: a002675. https://doi.org/10.1101/cshperspect.a002675

Johnson, M., I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis et al., 2008 NCBI BLAST: a better web interface. Nucleic Acids Res. 36: W5–W9. https://doi.org/10.1093/nar/gkn201

Johnson, D. S., C. Cinnioglu, R. Ross, A. Filby, G. Gemelos et al., 2010 Comprehensive analysis of karyotypic mosaicism between trophectoderm and inner cell mass. Mol. Hum. Reprod. 16: 944–949. https://doi.org/10.1093/molehr/gaq062

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. https://doi.org/10.1038/nature10413

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. https://doi.org/10.1126/science.1245503

Lanigan, T. M., H. C. Kopera, and T. L. Saunders, 2020 Principles of genetic engineering. Genes (Basel) 11: 291. https://doi.org/10.3390/genes11030291

Le Gall, J., M. Nizon, O. Pichon, J. Andrieux, S. Audebert-Bellanger et al., 2017 Sex chromosome aneuploidies and copy-number variants: a further explanation for neurodevelopmental prognosis variability? Eur. J. Hum. Genet. 25: 930–934. https://doi.org/10.1038/ejhg.2017.93

McCoy, R. C., 2017 Mosaicism in preimplantation human embryos: when chromosomal abnormalities are the norm. Trends Genet. 33: 448–463. https://doi.org/10.1016/j.tig.2017.04.001

McCray, Jr., P. B., L. Pewe, C. Wohlford-Lenane, M. Hickey, L. Manzel et al., 2007 Lethal infection of K18-hACE2 mice infected with severe acute respiratory syndrome coronavirus. J. Virol. 81: 813–821. https://doi.org/10.1128/JVI.02012-06

Morgan, A. P., and F. Pardo-Manuel de Villena, 2017 Sequence and structural diversity of mouse Y chromosomes. Mol. Biol. Evol. 34: 3186–3204. https://doi.org/10.1093/molbev/msx250

Morgan, A. P., and C. E. Welsh, 2015 Informatics resources for the Collaborative Cross and related mouse populations. Mamm. Genome 26: 521–539. https://doi.org/10.1007/s00335-015-9581-z

Morgan, A. P., C.-P. Fu, C.-Y. Kao, C. E. Welsh, J. P. Didion et al., 2015 The mouse universal genotyping array: from substrains to subspecies. G3 (Bethesda) 6: 263–279. https://doi.org/10.1534/g3.115.022087

Morgan, A. P., T. A. Bell, J. J. Crowley, and F. Pardo-Manuel de Villena, 2019 Instability of the pseudoautosomal boundary in house mice. Genetics 212: 469–487. https://doi.org/10.1534/genetics.119.302232

Mortazavi M., Y. Ren, S. Saini, D. Antaki, C. St. Pierre, et al., 2020 Importance of polymorphic SNPs, short tandem repeats and structural variants for differential gene expression among inbred C57BL/6 and C57BL/10 substrains. bioRXiv. doi: 10.1101/2020.03.16.993683 (Preprint posted March 18, 2020). https://doi.org/10.1101/2020/03/16/993683

Mulligan, M. K., X. Wang, A. L. Adler, K. Mozhui, L. Lu et al., 2012 Complex control of GABA(A) receptor subunit mRNA expression: variation, covariation, and genetic regulation. PLoS One 7: e34586. https://doi.org/10.1371/journal.pone.0034586

Peirce, J. L., L. Lu, J. Gu, L. M. Silver, and R. W. Williams, 2004 A new set of BXD recombinant inbred lines from advanced intercross populations in mice. BMC Genet. 5: 7. https://doi.org/10.1186/1471-2156-5-7

Rosshart, S. P., B. G. Vassallo, D. Angeletti, D. S. Hutchinson, A. P. Morgan et al., 2017 Wild mouse gut microbiota promotes host fitness and improves disease resistance. Cell 171: 1015–1028.e13. https://doi.org/10.1016/j.cell.2017.09.016

Sarsani, V. K., N. Raghupathy, I. T. Fiddes, J. Armstrong, F. Thibaud-Nissen et al., 2019 The genome of C57bl/6J "eve", the mother of the laboratory mouse genome reference strain. G3 (Bethesda) 9: 1795–1805. https://doi.org/10.1534/g3.119.400071

Scalzo, A. A., and Y. M. Yokoyama, 2008 Cmv1 and natural killer cell responses to murine cytomegalovirus infection. Curr. Top. Microbiol. Immunol. 321: 101–122. https://doi.org/10.1007/978-3-540–75203–5_5

Searle, J. B., and R. M. Jones, 2002 Sex chromosome aneuploidy in wild small mammals. Cytogenet. Genome Res. 96: 239–243. https://doi.org/10.1159/000063017

Shorter, J. R., F. Odet, D. L. Aylor, W. Pan, C.-Y. Kao et al., 2017 Male infertility is responsible for nearly half of the extinction observed in the mouse collaborative cross. Genetics 206: 557–572. https://doi.org/10.1534/genetics.116.199596

Shorter, J. R., M. L. Najarian, T. A. Bell, M. Blanchard, M. T. Ferris et al., 2019 Whole genome sequencing and progress toward full inbreeding of the mouse collaborative cross population. G3 (Bethesda) 9: 1303–1311. https://doi.org/10.1534/g3.119.400039

Smith, C. M., M. K. Proulx, R. Lai, M. C. Kiritsy, T. A. Bell et al., 2019 Functionally overlapping variants control tuberculosis

susceptibility in collaborative cross mice. mBio 10: e02791-19. https://doi.org/10.1128/mBio.02791-19

Srivastava, A., A. P. Morgan, M. L. Najarian, V. K. Sarsani, J. S. Sigmon et al., 2017   Genomes of the mouse collaborative cross. Genetics 206: 537–556. https://doi.org/10.1534/genetics.116.198838

Steemers, F. J., W. Chang, G. Lee, D. L. Barker, R. Shen et al., 2006   Whole-genome genotyping with the single-base extension assay. Nat. Methods 3: 31–33. https://doi.org/10.1038/nmeth842

Taylor B. A., H. J. Heiniger, and H. Meier, 1973   Genetic analysis of resistance to cadmium-induced testicular damage in mice. Proc. Soc. Exp. Biol. Med. 143: 629–633. https://doi.org/10.3181/00379727-143-37380

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 Off. J. Int. Mamm. Genome Soc. 10: 335–348. https://doi.org/10.1007/s003359900998

Treger, R. S., S. D. Pope, Y. Kong, M. Tokuyama, M. Taura et al., 2019   The lupus susceptibility locus Sgp3 encodes the suppressor of endogenous retrovirus expression SNERV. Immunity 50: 334–347.e9. https://doi.org/10.1016/j.immuni.2018.12.022

Veale, A. J., J. C. Russell, and C. M. King, 2018   The genomic ancestry, landscape genetics and invasion history of introduced mice in New Zealand. R. Soc. Open Sci. 5: 170879. https://doi.org/10.1098/rsos.170879

Yang, H., T. A. Bell, G. A. Churchill, and F. Pardo-Manuel de Villena, 2007   On the subspecific origin of the laboratory mouse. Nat. Genet. 39: 1100–1107. https://doi.org/10.1038/ng2087

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. https://doi.org/10.1038/nmeth.1359

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. Nat. Genet. 43: 648–655. https://doi.org/10.1038/ng.847

Yu, W., S. F. Hill, J. Xenakis, F. Pardo-Manuel de Villena, J. L. Wagnon et al., 2020   Gabra2 is a genetic modifier of Scn8a encephalopathy in the mouse. Epilepsia. https://doi.org/10.1111/epi.16741

Communicating editor: D. Greenstein