

1 **Content and performance of the MiniMUGA genotyping array, a new tool to**
2 **improve rigor and reproducibility in mouse research**

3
4
5 John Sebastian Sigmon^{*,1}, Matthew W Blanchard^{*,2,3}, Ralph S Baric⁴, Timothy A Bell², Jennifer
6 Brennan³, Gudrun A Brockmann⁵, A Wesley Burks⁶, J Mauro Calabrese^{7,8}, Kathleen M Caron⁹,
7 Richard E Cheney⁹, Dominic Ciavatta², Frank Conlon¹⁰, David B Darr⁸, James Faber⁹, Craig
8 Franklin¹¹, Timothy R Gershon¹², Lisa Gralinski⁴, Bin Gu⁹, Christiann H Gaines², Robert S Hagan¹³,
9 Ernest G Heimsath^{8,9}, Mark T Heise², Pablo Hock², Folami Iderabdullah^{2,14}, J. Charles
10 Jennette¹⁵, Tal Kafri¹⁶, Anwica Kashfeen¹, Samir Kelada², Mike Kulis⁶, Vivek Kumar¹⁷, Colton
11 Linnertz², Alessandra Livraghi-Butrico¹⁸, Kent Lloyd¹⁹, Richard Loeser^{20,21}, Cathleen Lutz¹⁷,
12 Rachel M Lynch², Terry Magnuson^{2,3,8}, Glenn K Matsushima¹⁶, Rachel McMullan², Darla Miller²,
13 Karen L Mohlke², Sheryl S Moy^{22,23}, Caroline Murphy², Maya Najarian¹, Lori O'Brien⁹, Abraham
14 A Palmer²⁴, Benjamin D Philpot^{9,25}, Scott Randell⁹, Laura Reinholdt¹⁷, Yuyu Ren²⁴, Steve
15 Rockwood¹⁷, Allison R Rogala^{15,26}, Avani Saraswatula², Christopher M Sasseti²⁷, Jonathan C
16 Schisler⁷, Sarah A Schoenrock², Ginger Shaw², John R Shorter², Clare M Smith²⁷, Celine L St.
17 Pierre²⁴, Lisa M Tarantino^{2,28}, David W Threadgill²⁹, William Valdar², Barbara J Vilen¹⁶, Keegan
18 Wardwell¹⁷, Jason K Whitmire², Lucy Williams², Mark Zylka⁹, Martin T Ferris², Leonard
19 McMillan¹, Fernando Pardo-Manuel de Villena^{2,3,8}

20
21 1, Department of Computer Science, UNC; 2, Department of Genetics, UNC; 3, MMRRC at UNC;
22 4, Department of Epidemiology, Gillings School of Public Health, UNC; 5, Humbolt-University
23 Berlin; 6, Department of Pediatrics, UNC; 7, Department of Pharmacology, UNC; 8, Lineberger
24 Comprehensive Cancer Center, UNC; 9, Department of Cell Biology and Physiology, UNC; 10,
25 Department of Biology, UNC; 11, Department of Veterinary Pathobiology, University of
26 Missouri; 12, Department of Neurology, UNC; 13, Division of Pulmonary Diseases and Critical
27 Care Medicine, UNC; 14, Department of Nutrition, Gillings School of Public Health, UNC; 15,
28 Department of Pathology and Laboratory Medicine, UNC; 16, Department of Microbiology and
29 Immunology, UNC; 17, the Jackson Laboratory; 18, Marsico Lung Institute, UNC; 19, University
30 of California Davis; 20, Division of Rheumatology, Allergy and Immunology, UNC; 21, Thurston
31 Arthritis Center, UNC; 22, Department of Psychiatry, UNC; 23, Carolina Center for
32 Developmental Disabilities, UNC; 24, University of California at San Diego; 25, UNC
33 Neuroscience Center; 26, Division of Comparative Medicine, UNC; 27 Department of
34 Microbiology and Physiological Systems, University of Massachusetts Medical School; 28,
35 Division of Pharmacotherapy and Experimental Therapeutics, Eshelman School of Pharmacy,
36 UNC; 29, Department of Molecular and Cellular Medicine, and Department of Biochemistry and
37 Biophysics, Texas A&M University

38
39 *, these authors have contributed equally to this work

40
41 **Corresponding authors:** Fernando Pardo-Manuel de Villena
42 Leonard McMillan
43 Martin T Ferris

44

45 **Key words:** genetic QC, genetic background, substrains, chromosomal sex, genetic constructs,
46 diagnostic SNPs

47

48 **Short Title:** A Platform for Genetic QC for the Mouse

49 **Abstract**

50 The laboratory mouse is the most widely used animal model for biomedical research, due in
51 part to its well annotated genome, wealth of genetic resources and the ability to precisely
52 manipulate its genome. Despite the importance of genetics for mouse research, genetic quality
53 control (QC) is not standardized, in part due to the lack of cost effective, informative and robust
54 platforms. Genotyping arrays are standard tools for mouse research and remain an attractive
55 alternative even in the era of high-throughput whole genome sequencing. Here we describe the
56 content and performance of a new Mouse Universal Genotyping Array (MUGA). MiniMUGA, an
57 array-based genetic QC platform with over 11,000 probes. In addition to robust discrimination
58 between most classical and wild-derived laboratory strains, MiniMUGA was designed to contain
59 features not available in other platforms: 1) chromosomal sex determination, 2) discrimination
60 between substrains from multiple commercial vendors, 3) diagnostic SNPs for popular
61 laboratory strains, 4) detection of constructs used in genetically engineered mice, and 5) an
62 easy to interpret report summarizing these results. In-depth annotation of all probes should
63 facilitate custom analyses by individual researchers. To determine the performance of
64 MiniMUGA we genotyped 6,899 samples from a wide variety of genetic backgrounds. The
65 performance of MiniMUGA compares favorably with three previous iterations of the MUGA
66 family of arrays both in discrimination capabilities and robustness. We have generated publicly
67 available consensus genotypes for 241 inbred strains including classical, wild-derived and
68 recombinant inbred lines. Here we also report the detection of a substantial number of XO and
69 XYY individuals across a variety of sample types, the extension of the utility of reduced
70 complexity crosses to genetic backgrounds other than C57BL/6, and the robust detection of 17
71 genetic constructs. There is preliminary but striking evidence that the array can be used to
72 identify both partial sex chromosome duplication and mosaicism, and that diagnostic SNPs can
73 be used to determine how long inbred mice have been bred independently from the main stock
74 for a significant action of the genotyped inbred samples. We conclude that MiniMUGA is a
75 valuable platform for genetic QC and important new tool to the increase rigor and
76 reproducibility of mouse research.

77

78

79 **INTRODUCTION**

80 The laboratory mouse is among the most popular and extensively used platforms for
81 biomedical research. For example, in 2018 over 82,000 scientific manuscripts available in
82 PubMed included the word “mouse” in the abstract. The laboratory mouse is such an attractive
83 model due to the existence of hundreds of inbred strains and outbred lines designed to address
84 specific questions, as well as the ability to edit the mouse genome; originally by homologous
85 recombination and now with more efficient and simple techniques such as CRISPR (Dong *et al.*
86 2019; Ayabe *et al.* 2019). The centrality of genetics in mouse-enabled research begs the
87 question of how genetic quality control (QC) is performed in these experiments.

88 We have a long track record of developing genotyping arrays for the laboratory mouse, from
89 the Mouse Diversity Array (Yang *et al.* 2009) to the previous versions versions of the Mouse
90 Universal Genotyping Array (MUGA,(Morgan *et al.* 2015)). These tools were originally designed
91 for the genetic characterization of two popular genetic reference populations, the Collaborative
92 Cross (CC) and the Diversity Outbred (DO), and then used for many other laboratory strains as
93 well as wild mice (Yang *et al.* 2011; Collaborative Cross Consortium 2012; Carbonetto *et al.*
94 2014; Arends *et al.* 2016; Didion *et al.* 2016; Shorter *et al.* 2017; Srivastava *et al.* 2017; Rosshart
95 *et al.* 2017; Veale *et al.* 2018). Efforts to extend the use of MUGA to characterize copy number
96 variation and genetic constructs were met with limited success (Morgan *et al.* 2015). In
97 conclusion, current genotyping tools are suboptimal for genetic QC and for new experimental
98 designs aimed at facilitating the rapid identification of causal genetic variants in mouse crosses.

99 An improved genotyping platform would ideally be able to provide reliable information about
100 the sex, genetic background and presence of genetic constructs in a given sample in a robust
101 and cost-effective manner. The ability to discriminate between most genetic backgrounds is
102 critical for genetic QC. The success of a new genotyping platform depends on how it compares
103 to other more comprehensive solutions such as whole genome sequence (WGS) in terms of
104 cost and ease involved in generating, analyzing, and interpreting the data. This is important
105 because many analyses require more sophisticated approaches and skills that are beyond many
106 users of laboratory mice. In addition, a new platform is needed to extend the success of
107 reduced complexity crosses (RCC) beyond the C57BL/6J – C57BL/6NJ pair of strains (Kumar *et*
108 *al.* 2013; Babbs *et al.* 2019). RCC are predicated on the idea that if a genetically driven
109 phenotype is variable between a pair of closely related laboratory substrains, then QTL
110 mapping combined with a complete catalog of the few thousand variants that differ among
111 these substrains can lead to the rapid identification of the candidate causal variants (Kumar *et*
112 *al.* 2013). This addresses one of the major limitations of standard mouse crosses, namely the
113 cost in time and resources to move from QTL to quantitative trait variants (QTV). Genetic
114 mapping in experimental F2 populations requires assigning every genomic region to one of
115 three diplotypes based on their genotypes at segregating SNPs or other variants. The difficulty
116 in RCC is two fold: first, genetic variants are unknown because WGS is not publicly available for
117 most substrains; second, these variants are so rare (5-20K genome wide or one variant per 100
118 to 500 kb) that low pass WGS will miss the majority of them, complicating the analysis. In other
119 words, the feature that makes RCC attractive for rapid QTV identification also makes it very
120 difficult to implement.

121 To address these issues we created a fourth iteration of the MUGA family of arrays that we call
122 MiniMUGA. The central considerations for the design were to reduce genotyping costs, provide
123 broad discrimination between most inbred strains, support genetic mapping in dozens of
124 different RCCs involving multiple substrains available from commercial vendors, robustly
125 determine chromosomal sex, and reliably detect presence of popular genetic constructs.
126 MiniMUGA fulfills all these criteria and facilitates simple, uniform and cost effective standard
127 genetic QC, as well as serving the mouse community at large by providing a new tool for genetic
128 studies.

129

130 MATERIALS AND METHODS

131 *Reference samples*

132 A diverse panel of 6,899 samples was used for calibrating and evaluating the performance of
133 the array. The type of sample is provided in **Table 1**. To test the performance of each
134 individual marker, provide reliable consensus genotypes and asses diagnostic markers, several
135 biological and/or technical replicates for many inbred strains and F1 hybrids were included.
136 **Supplementary Table 1** provides comprehensive information about each of these samples
137 including sample name, type, whether it was genotyped in the initial or final version of the
138 array, whether the sample was used to determine consensus genotypes or thresholds for
139 chromosomal sex, chromosomal sex, basic QC metrics and values used to determine the
140 presence of 17 constructs. A complete description of the information provided in
141 **Supplementary Table 1** is available in the table legend.

142 DNA stocks for classical inbred strains were purchased from The Jackson Laboratory or provided
143 by the authors. DNA from most other samples was prepared from tail clips or spleens using the
144 DNeasy Blood & Tissue Kit (catalog no. 69506; Qiagen, Hilden, Germany).

145 *Microarray platform*

146 MiniMUGA is implemented on the Illumina Infinium HD platform (Steemers *et al.* 2006).
147 Invariable oligonucleotide probes 50 bp in length are conjugated to silica beads that are then
148 addressed to wells on a chip. Sample DNA is hybridized to the oligonucleotide probes and a
149 single-basepair templated-extension reaction is performed with fluorescently labeled
150 nucleotides. The relative signal intensity from alternate fluorophores at the target nucleotide is
151 processed into a discrete genotype call (AA, AB, BB) using the Illumina BeadStudio software.
152 Although the two-color Infinium readout is optimized for genotyping biallelic SNPs, both total
153 and relative signal intensity can also be informative for copy-number variation and construct
154 detection.

155 *Probe design*

156 Of the 11,125 markers present in the array, 10,819 (97.2%) are probes designed for biallelic
157 SNPs and the remaining 306 markers (2.6%) are probes designed to test the presence of genetic
158 constructs (**Supplementary Table 2**). Nucleotides are labeled such that only one silica bead is
159 required to genotype most SNPs, except the cases of [A/T] and [C/G] SNPs, which require two
160 beads. In order to maximize information content, target SNPs were biased toward single-bead

161 SNPs (mostly transitions). There are 10,721 single-bead assays and 404 two bead assays. The
162 transition:transversion ratio in SNPs (excluding constructs) is 3:1.

163 *Array hybridization and genotype calling*

164 Approximately 1.5 μ g genomic DNA per sample was shipped to Neogen Inc. (Lincoln, NE) for
165 array hybridization. Genotypes were called jointly for all reference samples using the GenCall
166 algorithm implemented in the Illumina BeadStudio software.

167 *Probe Annotation*

168 Probe design and performance of individual assays was used to annotate the array.

169 **Supplementary Table 2** provides a rich set of annotations for each marker including: marker
170 name, chromosome position, strand, probe sequence, performance, rsID, diagnostic value,
171 thresholds for construct probes. A complete description of the information provided in
172 **Supplementary Table 2** is available in the table legend.

173 *Chromosomal sex determination*

174 We selected a set of 2,348 control samples (1,108 males and 1,240 females) with known X and
175 Y chromosome number as determined through standard phenotypical sexing, which was
176 supported by genotype analysis when expected heterozygosity on chromosome X was known.
177 For each control sample, we first normalized the intensity values at each X and Y chromosome
178 marker by dividing the intensity (r) by the sample's median autosomal intensity. These
179 autosome-normalized intensity values are used in all subsequent sex-determination
180 calculations.

181 The first step of chromosomal sex determination was to identify sex-linked markers that
182 provide a consistent estimate of sex chromosome number with minimal noise. We identified
183 269 X and 72 Y sex-informative markers as those for which the ranges of median normalized
184 intensity as defined by their standard deviations do not overlap between male and female
185 controls (**Supplemental Figure 1**). This information is provided in **Supplementary Table 2**.

186 Next, we established chromosomal sex intensity threshold values. For each sample, we plotted
187 the median of the normalized intensity values at the X informative markers on the x axis and
188 median of the normalized intensity values at the Y informative markers on the y axis (**Figure 1**).
189 Based on this plot we identified four clusters in sample intensity that correspond to XX, XY, XO,
190 and XXY chromosomal sex. We defined thresholds as the midpoint between the relevant
191 clusters. There is a single Y threshold value (0.3), separating samples with or without a Y
192 chromosome. We identified two independent X threshold values (0.77 and 0.69) depending on
193 whether the sample has a Y chromosome or not. These threshold values were used to classify
194 the chromosomal sex of experimental samples into four groups, XX, XY, XO, or XXY.

195 *Generation of consensus genotypes*

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

201 For each of 241 inbred strains (**Supplementary Table 3**), we genotyped from 1 to 19 samples
202 (average 3.2 per strain). Most inbred strains (179) were genotyped more than once. For 53
203 strains (mostly BXD recombinant inbred lines) we did not genotype a male animal and thus Y
204 chromosome genotypes are not provided for those strains. Over half of the strains (146) were
205 genotyped only in the initial version of the array, so final content genotypes are missing in
206 those strains. See **Supplementary Table 1** for details.

207 We generated consensus genotype calls at all 10,819 of the autosomal, X, pseudo-autosomal
208 region (PAR), and Y chromosome markers (biallelic SNPs). For each consensus strain, at each
209 marker, we recorded the genotype calls in all of the constituent samples and determined the
210 consistency among these calls. For strains with more than one sample, if all calls are consistent,
211 the consensus genotype is shown in upper case (A,T,C,G,H,N). Partially consistent calls are
212 those with a mix of one or more calls of a single nucleotide and one or more H and/or N calls.
213 Partially consistent calls are shown in lower case, as are calls for strains with a single
214 constituent sample. Inconsistent calls are those for which two distinct nucleotides calls are
215 observed. For standard markers, inconsistent genotypes within a strain are shown as N in the
216 consensus. For partially diagnostic SNPs the consensus call is the diagnostic allele shown in
217 lower case. For CC strains, inconsistent consensus genotypes are shown as H, as these markers
218 can be heterozygous in such samples. For mitochondria and Y chromosome markers, consensus
219 calls follow the same rules except H calls are treated as N. **Supplementary Table 4** provides a
220 list of rules for generating all possible consensus calls. **Supplementary Table 5** provides a listing
221 of the consensus genotypes.

222 *Informative SNPs between closely related substrains*

223 To increase the specificity of MiniMUGA as a tool for discriminating between closely related
224 inbred strains, we used public data from several other studies providing genotype or whole-
225 genome sequence information (Yang *et al.* 2009; Keane *et al.* 2011; Adams *et al.* 2015; Morgan
226 *et al.* 2015). Most importantly, we included SNP variants that are segregating between
227 substrains. These SNPs were identified by whole genome sequencing of 33 substrains
228 performed as part of two ongoing collaborations (contributed by either MTF, RSB and MTH, or
229 MTF and CMS; **Table 2**). Finally, we included 339 variants discriminating substrains of C57BL/10
230 (provided by AAP, YR and CSP). Some of the 5,171 GigaMUGA probes included to cover the
231 genome uniformly in classical and wild-derived inbred strains were also informative for
232 substrains.

233 *Probes for genetically engineered constructs*

234 We selected 36 constructs commonly used in genetic engineering in the mouse. For each
235 construct, we obtained full length sequence from either Addgene or GenBank. We ran a BLAST
236 search (Johnson *et al.* 2008) on these sequences to identify 2-5 additional sequences which (a)
237 had high BLAST scores, and (b) were annotated as containing the relevant construct gene we
238 were searching for (all sequence accession numbers are in **Supplementary Table 8**). For each
239 construct, sequences were then aligned using the EMBOSS Water algorithm from EMBL-EBI
240 (https://www.ebi.ac.uk/Tools/psa/emboss_water/). We identified conserved 50-mers within
241 these alignments followed by a single A in the forward strand, or followed by a single T in the
242 reverse strand. These sequences were submitted to the Illumina BeadStudio design pipeline,

243 with a pseudo-SNP (A/G or T/C). Probes which passed a quality score threshold of 0.7 were
244 included in the array. In total we created 306 probes for these constructs (range 3-18, median 8
245 probes/construct).

246
247 In order to validate these probes, we first eliminated probes which had high intensity signal in
248 the 580 negative control samples (standard inbred mouse strains and F1 hybrids between
249 them). Next, among the remaining probes, we identified those with significantly variable
250 intensity among the remaining 6,319 samples in this study. In particular, we confirmed that,
251 where available, positive controls had high signal intensity.

252 This process left 163 validated probes. We noticed that signal intensity of validated probes was
253 often positively correlated with other validated probes with the same, or related target
254 constructs. All validated probes were then subject to a second round of BLAST for final
255 identification of the targeted constructs and to provide a biological basis for grouping of highly
256 correlated probes. These alignments are provided in **Supplementary Figure 2**. In total these
257 163 probes mapped to 17 biologically distinct constructs (see **Table 3**). Probes tracking the
258 hCMV enhancer can divided into two groups based on the clustering.

259 Once we selected the final set of validated probes for a specific construct, we used the per-
260 sample distribution of the sum validated probe intensity to manually identify conservative
261 threshold values for the presence and absence of each construct. We used the negative and
262 positive controls to set initial thresholds and then used the distribution of values to identify
263 breaks and set the final thresholds such that we minimize the number of samples misclassified
264 as positive or negative.

265

266 *Additional sample quality metrics*

267 Most quality metrics for genotyping arrays are based on genotype calls. However, intensity
268 based analyses, such as chromosomal sex determination, assume quasi-normal distribution of
269 marker intensities in a given sample (**Supplementary Figure 3**). In our dataset some samples
270 had significantly skewed and idiosyncratic intensity distributions. Among these samples there is
271 an excess of sex chromosome aneuploidies as called by our algorithm, many of which are in fact
272 errors.

273 To identify samples with poor performance we first identified 200 random samples with no
274 chromosomal abnormalities and confirmed that they have quasi-normal intensity distribution in
275 aggregate. We then computed a Power Divergence statistic (pd_stat; equivalent to Pearson's
276 chi-squared goodness of fit statistic for each sample, comparing to that distribution.

277 **Supplementary Figure 4** shows the distribution of pd_stat values in our entire dataset. We
278 selected 3,230 as the threshold, such that in samples with higher values the reported
279 chromosomal sex could be incorrect. This warning is particularly true for samples reported to
280 have sex chromosome aneuploidy. The threshold also ensures that in samples from species
281 other than *Mus musculus*, chromosomal sex determination is treated with skepticism.

282 To determine whether a high pd_stat had an effect on the accuracy of genotyping calls we
283 selected four pairs of different F1 mice ((A/JxCAST/EiJ)F1_M15765; (CAST/EiJxA/J)F1_F002;

284 (CAST/EiJxNZO/HILtJ)F1_F0019; (CAST/EiJxNZO/HILtJ)F1_F022;
285 (NZO/HILtJxNOD/ShiLtJ)F1_F0042; (NZO/HILtJxNOD/ShiLtJ)F1_F0042;
286 (PWK/PhJxNZO/HILtJ)F1_F0019 and (PWK/PhJxNZO/HILtJ)F1_M0001) that cover a variety of
287 pd_stat comparisons (high/low, medium/medium, and low/low). For each pair we first
288 determined the pairwise consistency of the genotypes calls and then compared these
289 genotypes to predicted calls for the consensus reference inbred strains. Pairwise comparison
290 consistencies in the autosomes excluding N calls vary between 99.5% and 100%. Similarly, the
291 consistency with predicted genotypes is very high (99.5%-100%). We conclude that the pd_stat
292 is independent of genotype call quality.

293 *Data availability*

294 Genotype calls, hybridization intensity data and consensus genotypes for inbred strains (both
295 raw and processed) for 6,899 samples are available for download at the Dataverse (upon
296 acceptance flat files with the data will be posted).

297

298 **RESULTS**

299 **Sample set, reproducibility and array annotation**

300 To test the performance of the MiniMUGA array we genotyped 6,899 DNA samples from a wide
301 range of genetic backgrounds, ages and tissues (**Supplementary Table 1**). These samples
302 include many examples of inbred strains, F1 hybrids, experimental crosses and cell lines (**Table**
303 **1**). The array content was designed in two phases and thousands of samples were genotyped
304 to determine the marker performance, information content and to improve different aspects of
305 the proposed use of the array for genetic QC. In the initial array that contained 10,171 makers,
306 5,604 samples were genotyped. The second phase added 954 markers, with an additional
307 1,295 samples genotyped. This results in 6,300 samples that were genotyped once and 225
308 samples were genotyped two or more times, resulting in a total of 6,525 unique samples. The
309 599 replicates were used to estimate the reproducibility of the genotype data. Overall, $99.6 \pm$
310 0.4% of SNP genotype calls were consistent between technical replicates (range 95.9% to
311 100%). The consistency rate is similar for replicates run within and between versions of the
312 array. Samples with lower consistency rates include wild-derived samples from more distant
313 species and subspecies (SPRET/EiJ, SFM, SMZ, MSM/MsJ and JF1/Ms), lower quality samples,
314 and cell lines. Inconsistency was typically driven by a small minority of markers and by “no
315 calls” in one or few of the technical replicates.

316 Probe design and performance of individual assays was used to annotate the array.

317 **Supplementary Table 2** contains the following information: 1) Marker name; 2) Chromosome;
318 3) Position; 4) Strand; 5-6) Sequences for one and two bead probes; 7-8) Reference and
319 alternate allele at the SNP; 9) Tier; 10) rsID; 11) Diagnostic information; 12) Uniqueness; 13) X
320 chromosome markers used to determine the presence and number of X chromosomes; 14) Y
321 chromosomes markers used to determine the presence of a Y chromosome; 15) Markers added
322 in the second phase.

323 **Improved chromosomal sex determination reveals sex chromosome aneuploidy due to strain-
324 dependent paternal non-disjunction**

325 Typically, genetic determination of sex of a mouse sample has relied on detecting the presence
326 of a Y chromosome. This approach does not estimate X chromosome dosage and thus lacks the
327 ability to identify samples with common types of sex chromosome aneuploidies. In contrast,
328 MiniMUGA uses probe intensity to discriminate between normal chromosomal sexes (XX and
329 XY) and two types of sex chromosome aneuploidies, XO and XXY (**Supplementary Table 1**). The
330 methodology (Materials and Methods) relies on median autosome-normalized intensity at 269
331 X chromosome markers and 72 Y chromosome markers. This approach provides a robust
332 framework to discriminate between at least four types of chromosomal sex (**Figure 1**). Our set
333 of 6,899 samples was composed of 3,507 unique females (no Y chromosome present) and 3,018
334 unique males (Y chromosome present).

335 We initially identified 54 samples as potential XO and XXY. However, in eight XO females the
336 pattern of heterozygosity and recombination in the X chromosome (**Supplementary Table 6**)
337 demonstrates that these are, in fact, normal XX females with abnormal intensities. We
338 developed a new QC test (pd_stat, see Materials and Methods) to identify samples in which

339 chromosomal sex determination is not accurate. Once these eight samples were removed, 46
340 samples that had sex chromosome aneuploidies remained. To determine the rate of aneuploidy
341 we only considered unique samples (not replicates). This results in 45 aneuploid samples
342 among 6,525 total unique samples, an overall 0.7% rate. This rate is driven by a highly
343 significant excess (7X) of sex chromosome aneuploids among the cell lines. Notably all these
344 aneuploids are XO. Among live mice there were 36 unique aneuploids and the rate is 0.55%,
345 similar but higher than the reported rate in wild mice and in humans (Searle and Jones 2002;
346 Chesler *et al.* 2016; Le Gall *et al.* 2017). In this dataset, XO females are observed at significantly
347 higher frequency than XXY males ($p=0.02$; 25 XO females and 11 XXY males) (**Table 1**).

348 For 22 of the 45 unique samples with sex chromosome aneuploidies, both parents were known
349 and informative for the X chromosome. This information allowed us to potentially determine
350 the parental origin of the missing (in XO) or the extra (in XXY) X chromosome based on the
351 haplotype inherited and recombination patterns observed (**Supplementary Table 6; Figure 2**).
352 Overall, the parental origin can be determined unambiguously in 21 of these samples, and in all
353 but one sample (95%) the aneuploidy is due to sex chromosome non-disjunction in the paternal
354 germ line (**Figure 2**). Note that this applies to both XO and XXY samples. Given the paternal
355 origin of most sex chromosome aneuploidies, we investigated whether the type of sire had an
356 effect. We observed a highly significantly ($p<0.00001$) excess of aneuploids in the progeny of
357 (CC029/Unc x CC030/GeniUnc)F1 hybrid males than in all other sires. Out of 180 male progeny
358 of this cross, 5% of genotyped samples were aneuploids and both XO and XXY were observed (3
359 XO and 6 XXY mice, respectively). There was also evidence of an excess of sex chromosome
360 aneuploids in progeny of sires with CC011/Unc background (5 XO females, **Supplementary**
361 **Table 6**). We conclude that sex chromosome aneuploidy is relatively common in lab mice,
362 originates predominantly in the paternal germ line and depends on the sire genotype. In some
363 backgrounds aneuploidy rate is a factor of magnitude higher than in the general population.

364 **Detection of sex chromosome mosaicism**

365 There were eight samples (two classified as XX, three as XXY and three as XO) with abnormal
366 chromosome Y intensities (either too low or two high) and with low number of chromosome Y
367 genotype calls (**Figure 1**). Because this pattern suggested mosaicism we performed several
368 additional analyses. As a test case, we selected the tail-derived sample TL9348 (also named
369 Unknown, **Supplementary Tables 1 and 6**) because it was expected to be a F1 hybrid male
370 derived from a C57BL/6J and 129X1/SvJ outcross, has questionable genotype quality and low
371 pd_stat. Based on chromosome intensity this sample was classified as an XXY male with low
372 chromosome Y intensity. Inspection of the genotype calls on chromosome X reveals a
373 significant excess of N calls compared to the autosomes ($p<0.00001$, **Supplementary Table 6**).
374 Furthermore, the H calls are consistent with the expected contribution of the two parental
375 inbred strains but at only a fraction of expected sites. These results suggest that the mosaicism
376 is due to the loss of both the Y chromosome and one of the two X chromosomes in a fraction of
377 cells. To test this hypothesis, we plotted the intensity of X chromosome markers for three
378 types of controls, C57BL/6J and 129X1/SvJ samples and heterozygous females as well as for the
379 suspected mosaic XXY sample (**Figure 3**). The pattern shown in this figure explain the observed
380 mix of N calls, heterozygous calls and C57BL/6J calls in the XXY sample and confirms its mosaic
381 nature. It further demonstrates that the X chromosome lost is the 129X1/SvJ one. Finally, we

382 can estimate the fraction of cells with XXY and XO constitution using the distance of each marker
383 to their corresponding C57BL/6J and het counterparts. Based on the analysis, we estimate that
384 approximately half of cells are XXY and the other half XO, a result that is also consistent with
385 reduction in the Y chromosome intensity by half. Considered together, these results indicate
386 that the mosaicism occurred early during development, a common observation in embryo
387 mosaicism in humans (Johnson *et al.* 2010; Fragouli *et al.* 2011; McCoy 2017).

388 Among the remaining seven potential mosaics, one was a cell line and thus mosaicism of the
389 sex chromosomes is not unexpected. For the other six samples we performed a similar analysis
390 as the one described above. In all cases the two sets of calls were consistent and thus suggest
391 chromosome Y mosaicism. However only the two samples with 50 or more genotype calls have
392 strong support for such a conclusion. In the Discussion we expand this analysis and provide
393 some guidance for users of the array.

394 Strain specific chromosome Y duplications

395 Among XY males there was a distinct cluster of 64 male samples with higher normalized median
396 Y chromosome intensity (**Figure 1**). These samples include five inbred C3H/HeJ, two F1 hybrid
397 males with a C3H/HeJ chromosome Y (**Figure 4a**) and 52 males derived from a C3H/HeJ by
398 C3H/HeNTac F2 intercross. The plot of the normalized Y chromosome intensity in these males
399 and 81 additional males with Y chromosomes derived from other C3H/He substrains (**Figure**
400 **4a**), revealed a clear separation between males carrying a Y chromosome from C3H/HeJ and
401 males carrying C3H/HeNCrl, C3H/HeNHsd, C3H/HeNRj, C3H/HeNTac and C3H/HeOuJ Y
402 chromosomes. Males with the high intensity Y chromosome also include two transgenic strains
403 from The Jackson Laboratory, B6C3-Tg(APPswe,PSEN1dE9)85Dbo/Mmjjax and B6;C3-Tg(Prnp-
404 SNCA*A53T)83Vle/J. Both strains were developed and/or maintained in B6C3H background
405 (WEBSITE).

406 To determine the origin of the higher median intensity in males with a C3H/HeJ Y chromosome,
407 we plotted the normalized intensities at MiniMUGA markers located on that chromosome
408 (**Figure 4b**). Inspection of this figure indicates that 54 consecutive markers have distinctly
409 higher intensity and are flanked by markers with intensities that are undistinguishable from
410 males with other C3H/He Y chromosomes. These markers define a 2.9 Mb region located on
411 the short arm of the Y chromosome containing eight known genes *Eif2s3y*, *Uty*, *Dxd3y*, *Usp9y*,
412 *Zfy2*, *Sry* and *Rbmy*, and 12 gene models (**Figure 4b**). We conclude that C3H/He substrain
413 differences are due to an intrachromosomal duplication that arose and was fixed in the
414 C3H/HeJ lineage after the isolation of that substrain in 1952 (Akeson *et al.* 2006). There are five
415 additional non-C3H/He samples with high normalized median chromosome Y intensity, four
416 technical replicates from a single DBA/1OlaHsd male and a single *Ax1*^{+/−} congenic mouse on a
417 C57BL/6 background (**Figure 4a**). Each case represents a different, independent (different
418 haplotype and different boundaries, **Supplementary Figure 5**) and very recent duplication of
419 the Y chromosome. These duplications were segregating within a closed colony. Given that we
420 have identified three independent large segmental duplications of the Y chromosome among
421 3,018 unique males, we estimate the mutation rate at 1/1000, a relatively high rate. This is
422 consistent with the segmental duplications reported in wild mice (Morgan and Pardo-Manuel
423 de Villena 2017).

424 **An effective tool for genetic QC in laboratory inbred strains**

425 To determine the performance of MiniMUGA among inbred strains we genotyped 779 samples
426 representing 241 inbred strains including 86 classical inbred strains, 34 wild-derived inbred
427 strains, 49 BXD recombinant inbred lines and 72 CC strains (**Supplementary Table 3**). We
428 created consensus genotypes for each inbred strain using both biological and technical
429 replicates (see Materials and Methods). The use of replicates strengthen the conclusions that
430 can be made from our genetic analyses as they provide a simple but robust method to
431 determine the performance of each SNP in each strain (see Discussion) as well as determining
432 the dates when diagnostic alleles arise and potentially became fixed (see below). We note that
433 for the CC strains, which are incompletely inbred (Srivastava *et al.* 2017), our consensus calls
434 were based on a small number of samples. As such, these consensus may not completely
435 reflect the individual genotype of any CC animal from a specific strain. Future sampling of a
436 wider range of genotypes from CC mice throughout the history of the CC colony will assist in
437 more accurate consensus genotypes for these strains.

438 Using the consensus genotypes we determined the number of informative markers for pairwise
439 combinations of all inbred strains. **Figure 5** summarizes the results for 83 classical inbred
440 strains. Over 90% of comparisons have at least 1,280 informative autosomal markers and all
441 but 0.52% of pairwise comparisons have more than 40 informative autosomal markers (2.1
442 markers per autosome). These statistics are exceptional given the limited number of markers in
443 the array, the inclusion of a large number of diagnostic markers, and a substantial number of
444 construct markers. Although our focus is on classical inbred strains, we extended the analysis
445 to include 37 wild-derived strains. For all 2,924 combinations of classical and wild derived
446 strains, the informativeness is high (mean = 3,224, min = 1,649, max = 3,827). In marked
447 contrast, combinations between wild-derived strains have a much wider range of informative
448 SNPs (from 93 to 3,410) due to a significant fraction of combinations with few to moderate
449 number of informative SNPs. The pairs of strains with the lowest number of informative SNPs
450 include pairs of strain from a taxa other than *Mus musculus* (for example SPRET/EiJ, SMZ and
451 XBS) and pairs of strains that are known to have close phylogenetic relationships (TIRANO/EiJ
452 and ZALENDE/EiJ; and PWD and PWK/PhJ; (Yang *et al.* 2011)). We conclude that MiniMUGA is a
453 significant improvement for genotyping standard lab strains and experimental crosses derived
454 from them.

455 **Mitochondria**

456 MiniMUGA has 88 markers that track the mitochondrial genome, 82 of which segregate in our
457 set of 241 inbred strains. Based on these 82 markers, the inbred strains can be classified into
458 22 different haplogroups, 19 of which discriminate between *M. musculus* strains (**Figure 6a**).
459 Fifteen haplotypes represent *M. m. domesticus* (groups 1 to 15 in **Figure 6a**), and two
460 haplotypes represent *M. m. musculus* (16 and 17) and two *M. m. castaneus* (18 and 19). Three
461 haplotypes represent different species such as *M. spretus* and *M. macedonicus*.

462 In *M. musculus*, nine haplotypes are present in multiple inbred strains while 10 are found in a
463 single inbred strain. The most common haplotype is present in 158 inbred strains (including 49
464 BXD and 26 CC strains). This haplotype is found in many classical inbred strains including
465 C57BL6/J, BALB/cJ, A/J, C3H/HeJ, DBA/1J, DBA/2J and FVB/NJ. Unique haplotypes represent an

466 interesting mix of wild-derived strains (LEWES/EiJ, CALB/Rk, WMP/Pas, SF/CamEiJ, TIRANO/EiJ,
467 ZALENDE/EiJ, CIM) and DBA/2 substrains (DBA/2JOlaHsd and DBA/2NCrl). CC strains fall into six
468 common haplotypes, one shared by CC three founders A/J, C57BL/6J and NOD/ShiLtJ and five
469 haplotypes present in a single CC founder: PWK/PhJ, 129S1/SvlmJ, CAST/EiJ, NZO/HILtJ and
470 WSB/EiJ. Interestingly, SMZ, a wild-derived inbred strain of *M. spretus* origin, has a
471 mitochondrial haplotype that unambiguously cluster with *M. m. domesticus* (**Figure 6a**)
472 demonstrating a case of interspecific introgression.

473 **Chromosome Y**

474 MiniMUGA has 75 markers that track the Y chromosome, 57 of which segregate in our set of
475 189 inbred strains with at least one male genotyped. Based on these 57 markers, the inbred
476 strains can be classified into 18 different haplogroups, 16 of which are *M. musculus* (**Figure 6b**).
477 Only four haplotypes represent *M. m. domesticus*, two haplotypes represent *M. m. castaneus*
478 and 11 represent *M. m. musculus*. *M. spretus* and *M. macedonicus* are represented by a single
479 haplotype each. In *M. musculus*, all but one haplotype (CIM) are present in multiple inbred
480 strains. No single haplotype dominates in our collection of inbred strains (the most common is
481 present in 38 inbred strains). Interestingly, C57BL/6 substrains fall into three distinct
482 haplotypes. The ancestral haplotype is found in C57BL/6ByJ, C57BL/6NCrl, C57BL/6NHsd,
483 C57BL/6NJ, C57BL/6NRj and B6N-Tyr<c-Brd>/BrdCrCrl. This haplotype is present in other
484 classical inbred strains such as BALB/c, C57BL/10, C57BLKS/J, C57L/J and C58/J. The second
485 haplotype is present in C57BL/6JBomTac, C57BL/6JEiJ and C57BL/6JOlaHsd. Finally, C57BL/6J
486 has its own private haplotype shared with 10 CC strains. Each one of the eight founder strains
487 of the CC (A/J, C57BL/6J, 129S1/SvlmJ, NOD/ShiLtJ, NZO/HILtJ, CAST/EiJ, PWK/PhJ and WSB/EiJ)
488 has its own distinct haplotype.

489 **Diagnostic SNPs as tool for genetic QC and strain dating**

490 Almost 30% of the SNPs in MiniMUGA are diagnostic for a specific genetic background and were
491 selected from whole genome sequence of 45 classical inbred strains (**Table 2**). We define SNPs
492 as diagnostic when the minor allele is present only in a single classical inbred strain or in a set of
493 closely related substrains. The identification of these SNPs requires WGS from the
494 corresponding strain using the sequence of 12 publicly available strains (Keane *et al.* 2011;
495 Adams *et al.* 2015), 33 substrains that we sequenced and SNP data for the C57BL/10 strain
496 group (**Table 2**). We sequenced these substrains to develop MiniMUGA as well as the desire to
497 expand the number of strains amenable to RCC (MTF, unpub.). Although diagnostic SNPs have
498 low information content (i.e., most samples in a large set of genetically diverse mice will be
499 homozygous for the major allele) they fulfill these two critical missions. First, they increase the
500 specificity of the MiniMUGA array to identify the genetic background present in a sample. In
501 addition, they are essential to extend the power of genetic mapping in RCC beyond the
502 C57BL/6J-C57BL/6NJ paradigm (Kumar *et al.* 2013; Treger *et al.* 2019).

503 The 3,045 diagnostic SNPs can be divided into two classes based on whether they are diagnostic
504 for a specific substrain (i.e., BALB/cJBomTac or C3H/HeJ) or a strain group (i.e., BALB/c or
505 C3H/He). There are 2,408 SNPs that are diagnostic for one of 45 substrains and 637 SNPs
506 diagnostic for one of 10 strain groups (**Table 2**). A second classification divides diagnostic SNPs
507 into 2,910 fully diagnostic and 129 partially diagnostic SNPs. The difference between these two

508 classes is based on whether the diagnostic allele was fixed or was still segregating in the
509 samples used to determine the consensus genotypes of 46 classical inbred strains.

510 All diagnostic SNPs started as partially diagnostic SNPs and they highlight the often overlooked
511 fact that mutations arise in all stocks and some of them are fixed despite the best efforts to
512 reduce their frequency and impact. It should be theoretically possible to date when fully and
513 partially diagnostic SNPs arose and whether and when they became fixed in the main stock of an
514 inbred strain. This requires sampling a given substrain at known dates in the past in large
515 enough cohorts to make confident inferences, in other words genotype large cohorts isolated
516 from the main breeding line at known dates.

517 We have two such populations in our sample set, the BXD and the CC recombinant inbred lines
518 (RIL). In the former we determined whether diagnostic alleles for C57BL/6J and DBA/2J were
519 present in 49 BXD RILs. These RILs were generated in three different epochs: 22 of the
520 genotyped BXD lines belong to epoch I (E1, (Taylor *et al.* 1973)); four belong to epoch II (E2,
521 (Taylor *et al.* 1999)) and 23 belong to epoch III (E3, (Peirce *et al.* 2004)). We determined
522 whether the minor allele at diagnostic SNPs was observed first in epoch I, epoch II, or epoch III.
523 SNPs that were not observed in any of these epochs were grouped under the heading of post
524 E3. **Table 4a** summarizes these findings and further classifies the SNPs based on their diagnostic
525 information. We find similar patterns for C57BL/6J and DBA/2J diagnostic SNPs with epoch II
526 contributing the majority of diagnostic SNPs, and epochs III and IV contributing approximately
527 half each of the remaining SNPs. Epoch I SNPs are rare except for the DBA/2 strain group.
528 Finally, and as expected, all partial diagnostic SNPs for C57BL/6J belong to post E3.

529 The CC population offers another opportunity to annotate diagnostic SNPs as these RILs were
530 derived from mice from eight inbred strains in 2004 (Collaborative Cross Consortium 2012),
531 including four strains with diagnostics SNPs in MiniMUGA, C57BL/6J, A/J, 129S1/SvImJ and
532 NOD/ShiLtJ. We used 483 CC samples genotyped in the initial array to determine when these
533 334 diagnostic SNPs arose. We observe three types of patterns depending on the age of the
534 diagnostic allele: 1) the diagnostic allele is fixed in the CC population and thus the diagnostic
535 allele predates the start of the CC project in 2004; 2) the diagnostic allele is absent in the CC
536 population and thus the diagnostic allele arose after 2004; and 3) the allele is segregating in the
537 CC with some strains having fixed the diagnostic allele while it is absent in other CC strains.
538 **Table 4b** summarizes these findings.

539 In addition to determining when diagnostic SNPs arose, it is possible to estimate whether and
540 when they became fixed by examining the allele frequency at consecutive time points and for
541 consistency between populations. This is best exemplified for diagnostic SNPs of the C57BL/6J
542 substrains as we have two time points with substantial sampling, E3 with 23 BXD RIL and the
543 initiation of the CC with 72 CC RIL (note that only one eighth of them will have the C57BL/6J
544 haplotype at any given location and thus the real size of the population used to estimate
545 fixation is closer to 9). There are 75 SNPs that were labelled as fixed at E3 because they had
546 100% allele frequency in both BXD RIL and CC RILs with a C57BL6/J haplotype at the locus.
547 There are also 49 SNPs that were labelled as fixed at the start of the CC because they had 100%
548 allele frequency in CC RILs with a C57BL6/J haplotype at the locus. The remaining 26 diagnostic

549 SNPs were segregating or arose after the start of the CC project. The dates of origin and
550 fixation for diagnostic SNPs are provided in **Supplementary Table 2**.

551 The birth and fixation of diagnostic alleles can be used to determine the origin and breeding
552 history of a given sample of the appropriate background and thus estimate the expected level
553 of drift (see Discussion).

554 **Expansion of reduced representation crosses to a large number substrains**

555 We define RCC as crosses between substrains from a single laboratory strain that differ only at
556 mutations that arose after they were isolated and bred independently from the common inbred
557 stock. We tested the ability of MiniMUGA to efficiently cover the genome in 78 different RCC
558 between substrains for which we have consensus genotypes, whole genome sequences and for
559 which live mice are available from commercial vendors (see **Table 2**). We focus our analysis in
560 this group given that WGS of both substrains is required for rapid identification of causative
561 variant(s) (Kumar *et al.* 2013; Treger *et al.* 2019). We used the distance to the nearest
562 informative marker to estimate how well MiniMUGA covers the genome in a given RCC cross.
563 **Figure 7** summarizes these data and demonstrates that for 62 RCCs (82%) all of the genome is
564 covered by a linked marker and in 14 RCCs (18%) between 95% and 99.5% of the genome is
565 covered by a linked marker. Only in two RCCs (3%) there is a significant fraction genome that is
566 not covered by a linked marker. These two crosses are B6N-Tyr<c-Brd>/BrdCrCrl by
567 C57BL/6JOlalHsd and BALB/cByJ by BALB/cByJRj with 8% and 14% of the genome not covered,
568 respectively. An alternative test is the number of RCCs for which 95% of the genome is covered
569 by informative markers at 20cM (56 RCCs or 72%) and 40cM (72 RCCs or 92%) intervals. We
570 conclude that MiniMUGA provides a cost effective tool to extend RCC to substrains from the
571 129P, 129S, A, BALB/c, C57BL/6, C3H, DBA/1, DBA/2, FVB and NOD strains.

572 **Robust detection of common genetic constructs**

573 Given the broad usage of genetic editing technologies, a key design criterion of MiniMUGA was
574 the ability to detect frequently used genetic constructs. Utilizing our pipeline (low construct
575 probe intensity in classical inbred and F1 samples; variable intensity across the rest of our test
576 population), we positively identified samples containing 17 construct types (**Figure 8**).
577 Importantly, for eight of these constructs, our sample set included positive controls. These
578 positive controls showed robust detection of their relevant constructs. We detected further
579 positive samples from our set in both these eight constructs, as well as nine additional
580 construct classes. All such samples were in sample classes where constructs were plausible (e.g.
581 not wild-derived or CC samples), and there was high concordance for intensities among the
582 probes comprising the detection sets for each of these constructs.

583 For constructs with many probes (**Supplementary Figure 6, Supplementary Table 8**), we
584 noticed that samples we declared as positive could often have significant sample-to-sample
585 variation in their overall intensity (**Figure 8**). As described in the methods and **Supplementary**
586 **Table 8**, for some construct types our analysis suggested that some probes designed for
587 different constructs were in fact detecting conserved features among multiple construct types
588 (e.g. our 'g_FP' designation encompasses probes designed against green-, yellow-, and cyan-
589 fluorescent proteins). As such, it is possible that only a subset of our validated probes are

590 detecting any given sample's construct. Given our ability to positively identify construct classes
591 with as few as two probes, it is likely that even for constructs which have divergent sequences
592 from our designed sequences, or are targeting a more distantly related construct type, our
593 pipeline will flag samples. An alternative explanation for signal heterogeneity within a construct
594 class is due to within-sample heterogeneity. That is, samples either have variable copies of the
595 construct in question. Such observations might be more common in cell culture samples.
596 Alternatively, construct mosaicism in live animals may manifest as an intermediate signal for
597 given constructs.

598 As inferred from the above section, across these 17 constructs, we observed that our ability to
599 discriminate between negative and positive samples across these 17 constructs is strongly
600 correlated with the number of independent probes for that construct (**Supplementary Figure 2**,
601 **Figure 8**). As signal intensity is constrained by the dynamic range, our ability to definitively call
602 the presence of low probe number constructs is more uncertain. This uncertainty is especially
603 relevant where a given construct is genetically divergent from the construct sequences used to
604 define a given probe. Users are highly encouraged to consult the probe sequences when they
605 expect a given sample to contain a construct, but do not see support in the array itself.
606 Conversely, for constructs with many independent probes, positive support for a construct is
607 more conclusive, even if a given sample is not expected to contain any constructs.

608 Finally, we designed probes for 14 constructs, which universally failed in our pipeline. That is,
609 the intensity distributions between known negative (classical inbred strains from commercial
610 vendors and F1 hybrids) and experimental samples were not different. The easiest explanation
611 for these differences is that no samples within our set contained these constructs. Consistent
612 with this explanation is our *a priori* knowledge that no samples in our set could be defined as
613 known positives. In this case, probe-sets may in fact be diagnostic and individual users may
614 identify between sample intensity differences for these constructs. However, as the above
615 sections and methods caution, direct interpretation of single probes or probe-sets are
616 challenging without larger context. Alternatively, though less likely, is that our probe-sets will
617 fail regardless of construct presence. Definitive testing of construct-positive and construct-
618 negative samples for these probe-sets in the future will provide definitive answers to these.

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

620 The MiniMUGA Background Analysis Report (**Figure 9**) aims to provide users with essential
621 sample information derived from the genotyping array for every sample genotyped. The report
622 is designed to provide overall sample QC, as well as genetic background information for
623 classical inbred mouse strains, congenic, and transgenic mice. For samples outside of this scope
624 the report may be incomplete or provide misleading conclusions. Details of the thresholds and
625 algorithms for each section of the report are provided in the Materials and Methods section.

626 In addition to chromosomal sex and presence of constructs, the report provides a quantitative
627 and qualitative score for genotyping quality. Based on the number of N calls per sample of our
628 sample set we classified samples in one of four categories: samples with Excellent quality (0 to
629 91 N calls, represents 96.8% of samples); samples with Good quality (between 92 and 234 N
630 calls, 2% of samples), samples with Questionable quality (between 235 and 446 N calls, 0.9% of

631 samples) and samples with Poor quality (more than 447 N calls, 0.3% of samples). Only tier 1
632 and 2 markers were used in this analysis.

633 Regarding inbreeding status, the report assigns every sample to one of three categories: Inbred
634 (fewer than 61 H calls), close to inbred (between 61 and 280 H calls) and outbred (more than
635 280 H calls). These thresholds are based on the number of H calls observed in the autosomes of
636 172 samples of classical inbred strains and predicted heterozygosity in 3,655 *in silico* F₁ hybrid
637 mice (**Supplementary Figure 7**).

638 For genetic QC, the report provides two complementary analyses. One analysis determines the
639 primary and secondary background of a qualified sample based on the totality of its genotypes
640 (excluding the Y chromosome). The second returns the genetic backgrounds detected in a
641 sample based on the presence of the minor allele at diagnostic SNPs (see section on diagnostic
642 SNPs as tool for genetic QC and strain dating). The initial diagnostic analysis uses the presence
643 of minor alleles in the sample genotypes at identified diagnostic SNPs to identify which (if any)
644 of 46 substrains and/or 10 strain groups are present in the sample.

645 For the primary background analysis, the sample's genotype is compared to a set of 120
646 classical and wild-derived inbred reference strains (**Supplementary Table 3**) to identify the
647 strain that best explains the sample genotypes. If multiple substrains from the same strain
648 group have been detected via diagnostic alleles, or if there is an overrepresentation of a
649 particular diagnostic strain in the unexplained markers, the algorithm generates a composite
650 strain consensus that incorporates all substrains in that strain group and uses it in the primary
651 background analysis. The strain or combination of substrains that best matches the sample is
652 called the primary background for the sample. The report provides the number of homozygous
653 calls that are consistent or inconsistent with the primary background, as well as the number of
654 heterozygous calls in the sample. The primary background is always returned for samples in
655 which the primary background explains at least 99.8% of the sample genotype calls.

656 Once the primary background is determined, the algorithm tests whether at least 75% of the
657 markers inconsistent with the primary strain background or heterozygous (aka unexplained) are
658 spatially clustered. If they are not (<75% of markers spatially clustered) the algorithm will not
659 try to identify a secondary background. If at least 75% of the unexplained markers are
660 clustered, all strain(s) from the reference set that best explain at least half of the unexplained
661 calls are identified as secondary background(s). If the combination of primary and secondary
662 backgrounds explains at least 99.8% of the calls, the primary and secondary backgrounds are
663 reported. If it explains <99.8% then no genetic background is returned.

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

669 Note that this report is generated programmatically using the available reference inbred strains
670 (**Supplementary Table 3**). If the reported results are inconsistent with expectations, users need
671 to consider further analyses before reaching a final conclusion. All estimates and claims in the

672 report are heavily dependent on the quality of the sample and genotyping results. Less than
673 excellent genotyping quality will likely increase the likelihood of an incorrect conclusion.
674 Genotyping noise can lead to incorrect reporting and may be particularly misleading in samples
675 from standard commercial inbred strains. Fully inbred strains routinely have a small percentage
676 of spurious H calls. These do not represent true heterozygosity (see consensus of inbred
677 strains).

678 **Cell lines**

679 Cell lines can be subject to the same genetic QC as mice. We have previously reported that the
680 number of N calls is higher for cell lines that mixed tissues in other arrays (Didion *et al.* 2014).
681 There is some evidence of this in our dataset but it is inconclusive. We have already shown the
682 ability to detect sex chromosome aneuploidy in cell lines (**Figure 1**). Diagnostic SNPs can be used
683 to date cell lines in similar fashion with the added simplicity that cell lines are less susceptible to
684 change. Finally, cell line can be run the same Background Analysis Report pipeline discussed in
685 the previous section, some examples are provided in **Supplementary Figure 8**. The importance
686 of genetic QC in cell lines will grow in future given the increased emphasis on cell based
687 research.

688

689 **DISCUSSION**

690 **MiniMUGA as a tool for QC**

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

697 We urge users to pay particular attention to genotype quality, reported heterozygosity and
698 unexpected conclusions (i.e., sex, backgrounds and constructs detected). Genotype quality
699 depends on the sample quality, quantity and purity and on the actual genotyping process. Poor
700 sample quality can also be the byproduct of off target variants in the probes used for
701 genotyping and thus wild mouse samples and mice from related taxa are expected to have
702 lower apparent quality. Samples with poor quality will not be run through the report. Samples
703 with questionable quality may lead to incorrect conclusions. For samples of any quality the
704 total number of N calls should be carefully considered if unexpected results are reported. It is
705 also important to consider the pd_stat when evaluating the chromosomal sex determination.

706 Reported heterozygosity is sensitive to genotyping quality. A lower quality sample will typically
707 include more spurious heterozygous calls than an excellent quality sample of the same strain.
708 This leads to an incorrect estimate of the level of inbreeding in a given sample, and can be
709 particularly misleading in a fully inbred mouse of pure background. The thresholds used to
710 classify samples as inbred, close to inbred and outbred are somewhat arbitrary and reflect the
711 biases in SNP selection (overrepresentation of diagnostic SNPs for selected substrains) and the

712 highly variable range of diversity observed in F1 mice. We used the observed number of H calls
713 in known inbred samples and the predicted number of H calls among a large and varied set of
714 potential F1 hybrids to set our thresholds, but users should define the level of heterozygosity in
715 a specific experiment (**Supplementary Figure 7**). For example, mice generated in RCCs between
716 related substrains may have a very small number of H calls and thus will be misclassified as
717 more inbred than they really are. The report combines sample quality and heterozygosity in a
718 single figure for quick visual inspection. Note that the x and y axes are compressed in the high
719 value range to ensure that all samples, even those with very poor quality and/or high
720 heterozygosity, are shown. The precise location of a sample in the plot should help customers
721 contextualize their sample's quality and inbreeding when evaluating their results.

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

728 We anticipate that most users will use the Background Analysis Report to determine the genetic
729 background(s) present in a sample as well as their respective contributions. The identification
730 of the correct primary and secondary background is completely dependent on the pre-existing
731 set of reference strains (**Supplementary Table 3**). If a genotyped sample is derived from a
732 strain that is not part of this reference set, the reported results may be misleading or
733 completely incorrect. Users should consult the list of reference backgrounds. We expect the
734 number of reference backgrounds to increase over time, reducing the frequency and impact of
735 this problem. However, the current background detection pipeline is not appropriate for
736 recombinant inbred lines (RIL) such as the BXD and CC populations. By their very nature RIL
737 have mosaic genomes derived from two or more inbred strains (included in our panel) and thus
738 the background analysis will detect more than two inbred backgrounds (for CC strains) or
739 declare one of the parental strains as a secondary background in some specific cases. Users
740 interested in confirming or determining the identity of RIL can use our consensus genotypes to
741 do so.

742 An important caveat of the current primary and secondary background analysis is that the
743 approach is greedy, and all variants except those with H and N calls in the consensus are
744 considered. Because only a fraction of the SNPs are informative between a given pair of strains
745 (typically less than half, see **Figure 5** and **Supplementary Figure 7**), the algorithm always
746 overestimates the contribution of the primary background and underestimates the contribution
747 of the secondary background (**Figure 9**). As a general rule in congenic strains, the contribution
748 of the strain identified as the secondary background should be multiplied by at least 3. If the
749 exact contribution of either background is critical for the research question, the user should
750 reanalyze the data using only SNPs that discriminate between the two backgrounds.

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

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

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

765 Our construct probe design was focused on targeting conserved features of various genetic
766 engineering and/or in vitro constructs commonly used in mammalian genetics. We can split
767 these conserved probe-sets into two main classes: those for which we were able to detect
768 positive samples in this large cohort, and those for which we were not able to detect any
769 consistently positive signal/sample. Many of the probe-sets that are reported jointly as a single
770 construct type because the signals were highly correlated (e.g. the cyan, green and yellow
771 fluorescent protein probe-sets). Interested users can use the individual probe intensities to
772 refine the analysis.

773 Similarly, in the dataset used to define the performance of the array, we were unable to
774 identify samples positive for several individual probes and even some entire probe-sets
775 (**Supplemental Figure 6**). In some cases we excluded probes due to the fact that they work for
776 different subsets of samples than the included probes (see hTK_pr in **Supplementary Figure 6**).
777 In other cases, the excluded probes failed for an unknown reason and likely cannot be rescued
778 (iCRE in **Supplementary Figure 6**). Finally, addition of known positive controls may allow the
779 rescue of one or more of the 13 constructs targeted (e.g. ampicillin resistance **Supplementary**
780 **Figure 6**).

781 **MiniMUGA as a tool for discovery**

782 MiniMUGA was designed to support the research mission of geneticists, but the range of
783 applications will depend on the ingenuity of its users. In the results sections we explored three
784 areas in which MiniMUGA has high potential to complement existing resources and tools.

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

792 This type of analysis can also identify sex chromosome mosaicism (Johnson *et al.* 2010; Fragouli
793 *et al.* 2011; McCoy 2017) and large structural variants involving the sex chromosomes. In the
794 results section we have shown that mosaics are outliers from the four defined clusters

795 observed in the intensity based chromosome sex determination plot (**Figure 1**). Specifically,
796 they have abnormal Y chromosome intensities. These mosaics may also have an abnormally
797 high ratio of N calls in the X chromosome compared to the autosomes and chromosome X
798 marker intensity distributions biased towards one parent (**Figure 3**). The last analyses are only
799 possible in the presence of heterozygosity on the X chromosome.

800 In addition, MiniMUGA revealed a 6Mb *de novo* duplication of the distal chromosome X
801 (**Supplementary Figure 9**) in an F2 male. The size of this duplication is not large enough to
802 affect chromosomal sex determination and its discovery was due to the presence of 10
803 heterozygous calls clustered on distal X. These heterozygous calls occur at informative markers
804 between the two CC strains involved in the F2 cross and are embedded in a region of 26
805 consecutive markers with higher than expected intensity (**Supplementary Figure 9**).
806 Interestingly, the parental CC strains (CC029/Unc and CC030/GeniUnc) are the same for which a
807 10X increase in sex chromosome aneuploidy is observed. We concluded that this F2 male had a
808 sharply defined duplication of the distal X chromosome. These vignettes provide a potential
809 blueprint that can be extended to other chromosomes and structural variants. It also highlights
810 the importance of having a large set of well-defined genotyped controls, against which to
811 compare a given sample.

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

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

835 To test whether it is possible to use the annotated diagnostic SNPs to determine the age and
836 breeding history of a given sample or stock we selected 485 samples that were over inbred,

837 had over 99% identity to C57BL/6J and had no diagnostic alleles for any other substrain. The
838 analysis is based in the pattern of ancestral diagnostic SNPs that are classified as fixed in epoch
839 III (E3) or prior to the CC based. **Figure 10** shows three examples with different patterns. Panel
840 A shows a KO mouse from line created prior to epoch III (E3) and bred independently from the
841 C57BL/6J stock since at least 2004. The former conclusion is based the fact that we detect the
842 ancestral allele at 21 SNPs that were fixed prior to epoch III. The later is based in the
843 observation of ancestral alleles at 36 SNPs that we believe to be fixed by 2004 (21 and 15 from
844 E3 and CC, respectively) and that these markers are distributed across 14 chromosomes. Panel
845 B shows a transgenic mouse from a line created prior to the initiation of the CC (2004) and bred
846 independently from the C57BL/6J stock since them. Both conclusions are based the fact that
847 there are zero ancestral alleles at any of 75 diagnostic SNPs fixed by epoch III, the detection of
848 the ancestral allele at 18 SNPs that were fixed prior to the CC and that these markers are
849 distributed across 13 chromosomes. Finally, panel C shows a wild type C57BL/6J mouse derived
850 from the JAX colony after 2004. The conclusion is based in the lack of ancestral alleles at any of
851 124 fixed diagnostic SNPs and the presence of a derived allele at three SNPs that arose after the
852 CC. Notably our conclusions were consistent with the expectations from the owners of these
853 samples. However, these are fairly simple examples but more complex and more interesting
854 patterns are plentiful in our dataset. For example, four samples from a congenic inbred stock
855 show evidence of both an old stock and new refreshing of the genome in recent years
856 (**Supplementary Figure 11**). Specifically, the presence of ancestral alleles at many diagnostic
857 SNPs fixed prior to epoch III and the start of the CC speaks of mouse line generated and bred
858 independently for many years. On the other hand, heterozygosity at some of these markers as
859 well as the presence of post CC diagnostic alleles indicates that this line we refreshed by
860 backcrossing to C57BL/6J in recent years. Both conclusions are correct as this stock was
861 imported by Mark Heise at UNC in 2014 and backcrossed once or a few times to JAX mice
862 before being maintained by brother sister mating. In addition to improving the genetic QC, we
863 believe that this type of analysis may provide researchers with critical information to guide
864 both experimental design and data analysis. Most important is the ability to estimate the
865 amount of drift that has taken and thus the amount of genetic variation present in that line but
866 absent in the main stock. We expect that use of MiniMUGA and the continued and rapid
867 annotation of diagnostic SNPs not only for C57BL/6J but for all inbred substrains offers an
868 opportunity to significantly improve the rigor and reproducibility of mouse research.

869
870
871

872 **Acknowledgments**

873 This work was supported in part by U24HG010100 (to LM and FPMV); U42OD010924 (to TM);
874 U19AI100625 (to RSB, MTH, FPMV and MTF) and P01AI132130 (to CS, MTF and FPMV);
875 R01GM121806 (to JMC), P50DA039841 (to LT), R01MH100241 (to LT, WV, and FPMV),
876 5R01HL128119 and 5R01DK058702 (to TK), and U42 OD010921 (to CL and LR). The array was
877 used for authentication of key biological materials in the following grants: R01ES029925 and
878 P42ES031007 (to FPMV). The Systems Genetics Core Facility and Mutant Mouse Resource and
879 Research Center at the University of North Carolina provided in kind resources. MiniMUGA was
880 developed under a service contract to FPMV and LM from Neogen Inc., Lincoln, NE. None of the
881 authors have a financial relationship with Neogen Inc. apart from the service contract listed
882 above. The authors have no other conflict of interest to declare. We would like to acknowledge
883 Mohanish Deshmukh, Bev Koller, Helen Lazear, Lawrence E. Ostrowski, and Patrick Sullivan for
884 kindly providing some of the samples.

885

886 **REFERENCES**

887

888 Adams D. J., A. G. Doran, J. Lilue, and T. M. Keane, 2015 The Mouse Genomes Project: a

889 repository of inbred laboratory mouse strain genomes. *Mamm. Genome Off. J. Int.*

890 *Mamm. Genome Soc.* 26: 403–412. <https://doi.org/10.1007/s00335-015-9579-6>

891 Akeson E. C., L. R. Donahue, W. G. Beamer, K. L. Shultz, C. Ackert-Bicknell, *et al.*, 2006

892 Chromosomal inversion discovered in C3H/HeJ mice. *Genomics* 87: 311–313.

893 <https://doi.org/10.1016/j.ygeno.2005.09.022>

894 Arends D., S. Heise, S. Kärst, J. Trost, and G. A. Brockmann, 2016 Fine mapping a major obesity

895 locus (jObes1) using a Berlin Fat Mouse × B6N advanced intercross population. *Int. J.*

896 *Obes.* 2005 40: 1784–1788. <https://doi.org/10.1038/ijo.2016.150>

897 Ayabe S., K. Nakashima, and A. Yoshiki, 2019 Off- and on-target effects of genome editing in

898 mouse embryos. *J. Reprod. Dev.* 65: 1–5. <https://doi.org/10.1262/jrd.2018-128>

899 Babbs R. K., J. A. Beierle, Q. T. Ruan, J. C. Kelliher, M. M. Chen, *et al.*, 2019 *Cyfip1*

900 Haploinsufficiency Increases Compulsive-Like Behavior and Modulates Palatable Food

901 Intake in Mice: Dependence on *Cyfip2* Genetic Background, Parent-of Origin, and Sex.

902 *G3 Bethesda* Md 9: 3009–3022. <https://doi.org/10.1534/g3.119.400470>

903 Carbonetto P., R. Cheng, J. P. Gyekis, C. C. Parker, D. A. Blizzard, *et al.*, 2014 Discovery and

904 refinement of muscle weight QTLs in B6 × D2 advanced intercross mice. *Physiol.*

905 *Genomics* 46: 571–582. <https://doi.org/10.1152/physiolgenomics.00055.2014>

906 Chesler E. J., D. M. Gatti, A. P. Morgan, M. Strobel, L. Trepanier, *et al.*, 2016 Diversity Outbred

907 Mice at 21: Maintaining Allelic Variation in the Face of Selection. *G3 Bethesda* Md 6:

908 3893–3902. <https://doi.org/10.1534/g3.116.035527>

909 Collaborative Cross Consortium, 2012 The genome architecture of the Collaborative Cross

910 mouse genetic reference population. *Genetics* 190: 389–401.

911 <https://doi.org/10.1534/genetics.111.132639>

912 Didion J. P., R. J. Buus, Z. Naghashfar, D. W. Threadgill, H. C. Morse, *et al.*, 2014 SNP array

913 profiling of mouse cell lines identifies their strains of origin and reveals cross-

914 contamination and widespread aneuploidy. *BMC Genomics* 15: 847.

915 <https://doi.org/10.1186/1471-2164-15-847>

916 Didion J. P., A. P. Morgan, L. Yadgary, T. A. Bell, R. C. McMullan, *et al.*, 2016 R2d2 Drives Selfish

917 Sweeps in the House Mouse. *Mol. Biol. Evol.* 33: 1381–1395.

918 <https://doi.org/10.1093/molbev/msw036>

919 Dong Y., H. Li, L. Zhao, P. Koopman, F. Zhang, *et al.*, 2019 Genome-Wide Off-Target Analysis in
920 CRISPR-Cas9 Modified Mice and Their Offspring. *G3 Bethesda* Md 9: 3645–3651.
921 <https://doi.org/10.1534/g3.119.400503>

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

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

928 Johnson M., I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, *et al.*, 2008 NCBI BLAST: a
929 better web interface. *Nucleic Acids Res.* 36: W5-9. <https://doi.org/10.1093/nar/gkn201>

930 Johnson D. S., C. Cinnioglu, R. Ross, A. Filby, G. Gemelos, *et al.*, 2010 Comprehensive analysis of
931 karyotypic mosaicism between trophectoderm and inner cell mass. *Mol. Hum. Reprod.*
932 16: 944–949. <https://doi.org/10.1093/molehr/gaq062>

933 Keane T. M., L. Goodstadt, P. Danecek, M. A. White, K. Wong, *et al.*, 2011 Mouse genomic
934 variation and its effect on phenotypes and gene regulation. *Nature* 477: 289–294.
935 <https://doi.org/10.1038/nature10413>

936 Kumar V., K. Kim, C. Joseph, S. Kourrich, S.-H. Yoo, *et al.*, 2013 C57BL/6N mutation in
937 cytoplasmic FMRP interacting protein 2 regulates cocaine response. *Science* 342: 1508–
938 1512. <https://doi.org/10.1126/science.1245503>

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

943 McCoy R. C., 2017 Mosaicism in Preimplantation Human Embryos: When Chromosomal
944 Abnormalities Are the Norm. *Trends Genet. TIG* 33: 448–463.
945 <https://doi.org/10.1016/j.tig.2017.04.001>

946 Morgan A. P., C.-P. Fu, C.-Y. Kao, C. E. Welsh, J. P. Didion, *et al.*, 2015 The Mouse Universal
947 Genotyping Array: From Substrains to Subspecies. *G3 Bethesda Md* 6: 263–279.
948 <https://doi.org/10.1534/g3.115.022087>

949 Morgan A. P., and F. Pardo-Manuel de Villena, 2017 Sequence and Structural Diversity of
950 Mouse Y Chromosomes. *Mol. Biol. Evol.* 34: 3186–3204.
951 <https://doi.org/10.1093/molbev/msx250>

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

955 Rosshart S. P., B. G. Vassallo, D. Angeletti, D. S. Hutchinson, A. P. Morgan, *et al.*, 2017 Wild
956 Mouse Gut Microbiota Promotes Host Fitness and Improves Disease Resistance. *Cell*
957 171: 1015–1028.e13. <https://doi.org/10.1016/j.cell.2017.09.016>

958 Searle J. B., and R. M. Jones, 2002 Sex chromosome aneuploidy in wild small mammals.

959 Cytogenet. Genome Res. 96: 239–243. <https://doi.org/10.1159/000063017>

960 Shorter J. R., F. Odet, D. L. Aylor, W. Pan, C.-Y. Kao, *et al.*, 2017 Male Infertility Is Responsible for

961 Nearly Half of the Extinction Observed in the Mouse Collaborative Cross. Genetics 206:

962 557–572. <https://doi.org/10.1534/genetics.116.199596>

963 Srivastava A., A. P. Morgan, M. L. Najarian, V. K. Sarsani, J. S. Sigmon, *et al.*, 2017 Genomes of

964 the Mouse Collaborative Cross. Genetics 206: 537–556.

965 <https://doi.org/10.1534/genetics.116.198838>

966 Steemers F. J., W. Chang, G. Lee, D. L. Barker, R. Shen, *et al.*, 2006 Whole-genome genotyping

967 with the single-base extension assay. Nat. Methods 3: 31–33.

968 <https://doi.org/10.1038/nmeth842>

969 Taylor B. A., H. J. Heiniger, and H. Meier, 1973 Genetic analysis of resistance to cadmium-

970 induced testicular damage in mice. Proc. Soc. Exp. Biol. Med. Soc. Exp. Biol. Med. N. Y. N

971 143: 629–633. <https://doi.org/10.3181/00379727-143-37380>

972 Taylor B. A., C. Wnek, B. S. Kotlus, N. Roemer, T. MacTaggart, *et al.*, 1999 Genotyping new BXD

973 recombinant inbred mouse strains and comparison of BXD and consensus maps. Mamm.

974 Genome Off. J. Int. Mamm. Genome Soc. 10: 335–348.

975 <https://doi.org/10.1007/s003359900998>

976 Treger R. S., S. D. Pope, Y. Kong, M. Tokuyama, M. Taura, *et al.*, 2019 The Lupus Susceptibility

977 Locus Sgp3 Encodes the Suppressor of Endogenous Retrovirus Expression SNERV.

978 Immunity 50: 334–347.e9. <https://doi.org/10.1016/j.immuni.2018.12.022>

979 Veale A. J., J. C. Russell, and C. M. King, 2018 The genomic ancestry, landscape genetics and

980 invasion history of introduced mice in New Zealand. R. Soc. Open Sci. 5: 170879.

981 <https://doi.org/10.1098/rsos.170879>

982 Yang H., Y. Ding, L. N. Hutchins, J. Szatkiewicz, T. A. Bell, *et al.*, 2009 A customized and versatile

983 high-density genotyping array for the mouse. Nat. Methods 6: 663–666.

984 <https://doi.org/10.1038/nmeth.1359>

985 Yang H., J. R. Wang, J. P. Didion, R. J. Buus, T. A. Bell, *et al.*, 2011 Subspecific origin and

986 haplotype diversity in the laboratory mouse. Nat. Genet. 43: 648–655.

987 <https://doi.org/10.1038/ng.847>

988

989

990

991

992

993

994 **TABLES**

995 **Table 1**

996

Content	Chromosomal sex	Inbred	F1	CC	Cross	Unclassified	Cell lines	Total
Initial	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
Final	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
997	Total	597	247	537	3466	1868	184	6899
998								
999								
1000								

997

998

999

1000

1001 **Table 2**

Background	Strain group	Diagnostic Type	Full	Partial	WGS
129P2/OlaHsd	129P	substrain	25	0	Sanger
129P3/J	129P	substrain	54	0	UNC
129S1/SvImJ	129S	substrain	82	13	Sanger
129S2/SvHsd	129S	substrain	7	1	UNC
129S2/SvPasOrlRj	129S	substrain	36	0	UNC
129S4/SvJaeJ	129S	substrain	45	0	UNC
129S5/SvEvBrd	129S	substrain	12	0	Sanger
129S6/SvEvTac	129S	substrain	41	0	UNC
129T2/SvEmsJ	129T	substrain	38	0	UNC
129X1/SvJ	129X	substrain	39	0	UNC
A/J	A	substrain	58	7	Sanger
A/JCr	A	substrain	53	0	UNC
A/JOlA-Hsd	A	substrain	38	0	UNC
BALB/cAnNCrl	BALB/c	substrain	36	2	UNC
BALB/cAnNHsd	BALB/c	substrain	109	4	UNC
BALB/cByJ	BALB/c	substrain	3	4	UNC
BALB/cByJRj	BALB/c	substrain	19	0	UNC
BALB/cJ	BALB/c	substrain	103	3	Sanger
BALB/cJBomTac	BALB/c	substrain	47	0	UNC
C3H/HeJ	C3H/He	substrain	166	2	Sanger
C3H/HeNCrl	C3H/He	substrain	39	0	UNC
C3H/HeNHsd	C3H/He	substrain	39	1	UNC
C3H/HeNRj	C3H/He	substrain	42	0	UNC
C3H/HeNTac	C3H/He	substrain	45	14	UNC
C57BL/6J	C57BL/6	substrain	136	20	Reference
C57BL/6JBomTac	C57BL/6	substrain	41	2	UNC
C57BL/6JOlaHsd	C57BL/6	substrain	43	0	UNC
C57BL/6NJ	C57BL/6	substrain	37	7	Sanger
C57BL/6NRj	C57BL/6	substrain	20	0	UNC
B6N-Tyr<C-Brd>/BrdCrCrI	C57BL/6	substrain	21	10	UNC
DBA/1J	DBA/1	substrain	70	0	Sanger
DBA/1LacJ	DBA/1	substrain	77	2	UNC
DBA/1OlA-Hsd	DBA/2	substrain	32	0	UNC
DBA/2J	DBA/2	substrain	112	0	Sanger
DBA/2JOlaHsd	DBA/2	substrain	39	0	UNC
DBA/2JRj	DBA/2	substrain	30	0	UNC
DBA/2NCrl	DBA/2	substrain	85	14	UNC
DBA/2NTac	DBA/2	substrain	36	10	UNC
FVB/NCrl	FVB	substrain	47	0	UNC
FVB/NHsd	FVB	substrain	39	1	UNC
FVB/NJ	FVB	substrain	72	7	Sanger
FVB/NRj	FVB	substrain	47	0	UNC
FVB/NTac	FVB	substrain	37	0	UNC
NOD/MrkTac	NOD	substrain	33	0	UNC
NOD/ShiLtJ	NOD	substrain	51	3	Sanger
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	Abraham
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	Sanger
Subtotal			635	0	
TOTAL			2916	127	

1002

1003

1004

1005 **Table 3**

1006

Name	Abreviation	# of probes	# of distinct probes
"Greenish" Fluorescent Protein (EGFP, EYFP, ECFP)	g_FP	19	19
SV40 large T antigen	SV40	18	18
Cre recombinase	Cre	16	12
Tetracycline repressor protein	tTA	14	14
Diphtheria 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
Caspase 9	cas9	7	7
Blasticidin resistance	BlastR	6	4
Internal Ribosome Entry Site	IRES	6	6
hCMV enhancer ^{version a}	hCMV_a	5	4
"Redish" fluorescent protein (tdTomato, mCherry)	r_FP	6	6
Herpesvirus TK promoter	hTK_pr	2	2
Total		163	145

1007

1008

1009

1010

1011 **Table 4**

1012

A) Epoch	C57BL/6J		DBA/2J		C57BL/6J group	DBA/2J group	C57BL/6	DBA/2	Other
	Full	Partial	Full	Partial					
I	0	0	4	0	2	24	1	0	0
II	72	0	68	0	4(2)	0	0	0	1*
III	34	0	16	0	0	0	0	0	0
IV	30	20	24	0	0	0	0	0	0

B)	A/J		C57BL/6J		129S1/SvImJ		NOD/ShiLtJ	
	Full	Partial	Full	Partial	Full	Partial	Full	Partial
PreCC	47	0	116	7	75	6	34	0
DuringCC	8	3	16	7	2	4	2	0
PostCC	0	2	0	3	0	1	0	1

1013

1014

1015

1016

1017

1018 **FIGURES**

1019 **Figure 1.** Chromosomal sex determination in 6,899 samples. Each dot and cross represents
1020 one sample. The x value is the autosome normalized median sample intensity at 269 sex
1021 informative X chromosome markers, and the y value is the autosome normalized median
1022 sample intensity at 72 sex informative Y chromosome markers. The dot color denotes the
1023 assigned chromosomal sex: XX, red; XY, blue; XO, green; XXY, purple. Potential mosaic samples
1024 are shown in gray and known errors in yellow. Samples with normal pd_stat as shown as circles
1025 and samples with high pd_stat are shown as crosses.

1026 **Figure 2.** Paternal origin of most sex chromosome aneuploids. Only the sex chromosomes and
1027 the mitochondria are shown. The X chromosomes are shown as acrocentric, Y chromosomes as
1028 submetacentric and mitochondria as circles. The parents of two types of crosses (outcross or
1029 intercross) are shown at the top of the figure with the dam shown on the left and the sire on
1030 the right. The potential types of aneuploid progeny in each type of cross are shown with the
1031 parental origin below. The figure also shows the inferred parental origin of the aneuploidy and
1032 the actual number of those observed in our dataset.

1033 **Figure 3.** Complex sex chromosome mosaicism in an XXY male. a) shows the chromosomal sex
1034 and mitochondria complement of the parents and XXY progeny. b) was used to identify the sex
1035 chromosome aneuploidy (two X chromosomes and Y present) and as evidence of mosaicism for
1036 presence and absence of Y chromosome (low Y intensity). c) provides evidence of mosaicism for
1037 the X chromosome and identifies the paternal origin (129X1/SvJ) of the chromosome lost in
1038 some cells. d) The sex chromosome complement of the two types of cells present in this male
1039 are shown. Panels b and c were used also to estimate the fraction of each type of cells. (blue
1040 points denote C57BL/6J genotype calls, red points 129s1/SvImJ genotype calls. Panels a, c, d).

1041 **Figure 4.** Segmental chromosome Y duplications in laboratory strains. a) Normalized median Y
1042 chromosome intensity in selected samples with C3H/He, DBA/1 and C57BL/6 Y chromosomes.
1043 Samples with a C3H/HeJ Y chromosome are shown in orange while samples with any other
1044 C3H/He Y chromosome are shown in different shades of blue. b) Spatial distribution of
1045 normalized intensity at SNPs in the proximal end of the Y chromosome in the same C3H/He
1046 samples shown in the a panel. The range of intensities in samples with a C3H/HeJ Y
1047 chromosome are shown in orange while samples with any other type of C3H/He Y chromosome
1048 are shown in blue. Duplicated region is shown in red and transition regions with uncertain copy
1049 number are shown in pink. The bottom of the figure shows the location of the MiniMUGA
1050 markers and genes.

1051 **Figure 5.** Pairwise number of informative calls in classical inbred strains. Strains are ordered by
1052 similarity and colors represent the number of informative SNPs based on the consensus
1053 genotypes. Only homozygous base calls, at tier 1 and 2 markers, on the autosomes, X, and PAR
1054 are included.

1055 **Figure 6.** Haplotype diversity of the mitochondria (a) and chromosome Y (b). The trees are built
1056 based on the variation present in MiniMUGA and may not represent the real phylogenetic
1057 relationships. Colors denote the subspecies-specific origin of the haplotype in question: shades

1058 of blue represent *M. m. domesticus* haplotypes; shades of red represent *M. m. musculus*
1059 haplotypes; shades of green represent *M. m. castaneus* haplotypes.

1060 **Figure 7.** Percent of the genome covered by MiniMUGA in RCCs. The 78 RCCs are shown in
1061 ascending order independently for each one of the six analyses. Coverage was based on the
1062 linkage distance to the nearest informative marker in a given RCC cross.

1063 **Figure 8.** Detection of genetic constructs. For each construct, samples are classified as negative
1064 controls (left), experimental (center) and positive controls (right). The dot color denotes
1065 whether the sample is determined to be negative (blue), positive (red), or questionable (grey)
1066 for the respective construct. For each construct, the grey horizontal lines represent the
1067 thresholds for positive and negative results. Note for each construct, the Y-axis scale is
1068 different.

1069 **Figure 9.** Background Analysis Report for the sample B6.Cg-*Cdkn2a*^{tm3.1Nesh} *Tyr*^{c-2J} *Hr*^{hr}/Mmnc
1070 (named MMRRC_UNC_F38673). The genotype of this sample is of excellent quality. It is a close
1071 to inbred female that is a congenic mouse with C57BL/6J as a primary background, and with
1072 multiple regions of a 129S background. This sample is positive for a luciferase-family construct
1073 and negative for 16 other constructs.

1074 **Figure 10.** Age and breeding history of mouse samples with C57BL/6J background. a) Inbred
1075 *Baff*^{-/-} male in C57BL/6J background. b) Inbred transgenic and IFNgR1 female in C57BL/6J
1076 background. c) Inbred C57BL/6J male. Red bars denote the ancestral allele for diagnostic SNPs
1077 fixed at E3. Pink bars denote ancestral alleles for diagnostic SNPs fixed at the start of the CC.
1078 Light blue bars denote diagnostic alleles at diagnostic SNPs fixed at E3. Lighter blue bars denote
1079 diagnostic alleles at diagnostic SNPs fixed at start of CC. Grey bars denote ancestral alleles at
1080 post-CC diagnostic SNPs. Dark blue bars denote diagnostic alleles at post-CC diagnostic SNPs.
1081 Split bars denote heterozygous SNPs in a sample.

1082

1083

1084 **SUPPLEMENTARY MATERIAL LEGENDS**

1085 **Supplementary Table 1.** Samples included in this study. The table provides the following
1086 information:

1087 Serial ID.

1088 Sample name: name provided by the investigator.

1089 Type: inbred, F1, cell line, cross or unclassified.

1090 Content Type: initial or Final.

1091 Consensus strain: if a sample was used to build the consensus genotypes of one 241 inbred
1092 strain, that strain name is listed, if that sample was not used then zero.

1093 Chromosomal sex marker selection: TRUE for samples used in selecting sex informative
1094 markers. FALSE for samples not used.

1095 Chromosomal sex: XX, XY, XO, XXY or XX*. The latter group are XX samples misclassified as XO.

1096 Replicate: TRUE for technical replicates genotyped more than once. FALSE for samples
1097 genotyped only once.

1098 Replicate name: An unambiguous name for that group of replicate samples.

1099 X chromosome intensity: median normalized intensity of chromosome X sex-informative
1100 markers.

1101 Y chromosome intensity: median normalized intensity of chromosome Y sex-informative
1102 markers.

1103 Median autosomal intensity: median intensity of autosomal markers (i.e., normalization factor)

1104 H calls: Number of heterozygous calls for tier 1 and 2 markers (see below) in the autosomes and
1105 chromosome X.

1106 H call on chromosome X: Number of heterozygous calls for tier 1 and 2 markers (see below) on
1107 chromosome X.

1108 Autosomal N calls: Number of no calls for tier 1 and 2 markers (see below) in the autosomes.

1109 N calls on chromosome X: Number of no calls for tier 1 and 2 markers (see below) in the X
1110 chromosome.

1111 ks_stat: Kolmogorov-Smirnov goodness of fit test statistic of the sample's autosomal intensities
1112 against the autosomal intensity distribution of 200 random samples

1113 pd_stat: Pearson's chi-squared test statistic of the sample's autosomal intensities against the
1114 autosomal intensity distribution of 200 random samples.

1115 BlastR: Sum of the autosome-normalized xraw intensity at 6 markers used to declare the
1116 presence or absence of the construct Blasticidin resistance.

1117 Cas9: Sum of the autosome-normalized xraw intensity at 7 markers used to declare the
1118 presence or absence of the construct Cas9

1119 Cre: Sum of the autosome-normalized xraw intensity at 15 markers used to declare the
1120 presence or absence of the construct Cre recombinase

1121 DT: Sum of the autosome-normalized xraw intensity at 11 markers used to declare the presence
1122 or absence of the construct Diphteria toxin

1123 IRES: Sum of the autosome-normalized xraw intensity at 6 markers used to declare the
1124 presence or absence of the construct Internal Ribosome Entry Site

1125 Luc: Sum of the autosome-normalized xraw intensity at 10 markers used to declare the
1126 presence or absence of the construct Luciferase

1127 SV40: Sum of the autosome-normalized xraw intensity at 18 markers used to declare the
1128 presence or absence of the construct SV40 large T antigen

1129 bpA: Sum of the autosome-normalized xraw intensity at 8 markers used to declare the presence
1130 or absence of the construct Bovine growth hormone poly A signal sequence

1131 chlor: Sum of the autosome-normalized xraw intensity at 9 markers used to declare the
1132 presence or absence of the construct Chloramphenicol acetyltransferase

1133 g FP: Sum of autosome-normalized xraw intensity at 19 markers used to declare the presence
1134 or absence of the construct "Greenish" Fluorescent Protein (EGFP, EYFP, ECFP)

1135 hCMV a: Sum of the autosome-normalized xraw intensity at 6 markers used to declare the
1136 presence or absence of the construct hCMV enhancer version a.

1137 hCMV b: Sum of the autosome-normalized xraw intensity at 11 markers used to declare the
1138 presence or absence of the construct hCMV enhancer version b

1139 hTK pr: Sum of the autosome-normalized xraw intensity at 2 markers used to declare the
1140 presence or absence of the construct Herpesvirus TK promoter

1141 iCre: Sum of the autosome-normalized xraw intensity at 8 markers used to declare the presence
1142 or absence of the construct iCre recombinase

1143 r FP: Sum of the autosome-normalized xraw intensity at 5 markers used to declare the
1144 presence or absence of the construct "Reddish" fluorescent protein (tdTomato, mCherry)

1145 rtTA: Sum of the autosome-normalized xraw intensity at 8 markers used to declare the
1146 presence or absence of the construct Reverse improved tetracycline-controlled transactivator

1147 tTA: Sum of the autosome-normalized xraw intensity at 14 markers used to declare the
1148 presence or absence of the construct Tetracycline repressor protein

1149

1150

1151 **Supplementary Table 2.** Marker annotation. The table contains the following information:

1152 1) Marker name.

1153 2) Chromosome. The following types are allowed: 1-19, for the autosomes; X and Y for the sex
1154 chromosomes; PAR, for markers on the pseudoautosomal region; MT, for the mitochondria and
1155 0, for genetic constructs.

1156 3) Position in bases in build 38.

1157 4) Strand. +, indicating the probe sequence is found in the 5' to 3' order (on the forward strand)
1158 in the reference genome immediately preceding the variant. -, indicating that the reverse
1159 complement of the probe sequence is found in the 5' to 3' order (on the forward strand) in the
1160 reference genome, immediately following the variant and NA, when not available.

1161 5-6) Sequences A and B. Sequence A for one bead probes is the sequence of the marker probe
1162 without the SNP and for two bead probes, the sequence of the marker probe including the SNP.
1163 Sequence B: for one bead probes, not applicable; for two bead probes, the alternative
1164 sequence of the marker probe including the SNP.

1165 7-8) Reference Allele and Alternate allele. Columns denoting the genotype call for the reference
1166 and alternative alleles

1167 9) Tier. For biallelic SNP markers, tier was assigned based on observed genotype call types
1168 (homozygous reference, homozygous alternate, or heterozygous) at each marker across a set of
1169 3,878 samples used for array QC and validation. Tier 1 markers were those for which we
1170 observe all three call types. Tier 2 markers were those for which we observe two of the three
1171 call types. Tier 3 markers were those for which we observe only one call type. Tier 4 markers
1172 were those markers for which we observe no calls (N) in every sample. For construct markers,
1173 tier is assigned based on the capability of a marker to detect a given construct. Informative tier
1174 makers are those for which the marker has been validated to test for the presence or absence
1175 of a given construct based on intensity. Partially informative tier makers were those for that
1176 could potentially be used to test for the presence or absence of a given construct based on
1177 intensity. Those markers which have not been tested were assigned the tier “Not tested”.

1178 10) rsID.

1179 11) Diagnostic. Name of the construct, substrain or strain group that the marker is diagnostic
1180 for. In all other cases is empty.

1181 12) Diagnostic type. Substrain, strain group or construct.

1182 13) Diagnostic information: Abbreviated name of the construct, name of substrain or list of
1183 substrains in which we observed the diagnostic allele. In all other cases is empty.

1184 14) Partial diagnostic: 1, for diagnostic alleles that are not fixed. 0, in all other cases.

1185 15) Diagnostic allele. Whether the reference or alternative allele is the diagnostic

1186 16) Positive threshold. Threshold value to declare the presence of a given construct

1187 17) Negative threshold. Threshold value to declare the absence of a given construct.

1188 18) Uniqueness measured using Bowtie.

1189 19) X chromosome markers used to determine the presence and number of X chromosomes. 1,
1190 chromosome X markers used in sex chromosome determination. 0, in all other cases.

1191 20) Y chromosomes markers used to determine the presence of a Y chromosome. 1,
1192 chromosome Y markers used in sex chromosome determination. 0, in all other cases.

1193 21) Flags. SPIKE, markers added in the final iteration of the array. Empty in all other cases.

1194 22) Diagnostic Birth. The population where a diagnostic allele was first seen (E2, E3, E4 in the
1195 BXD, Pre-Cc, CC or Post-CC in the Collaborative Cross)

1196 23) Diagnostic Fixation. The population where a diagnostic allele is inferred to be fixed (E3, CC
1197 or segregating)

1198 **Supplementary Table 3.** List of inbred strains with consensus genotypes grouped into four
1199 classes: classical, wild-derived, CC and BXD.

1200 **Supplementary Table 4.** Examples of the rules for consensus genotypes calls. *, denotes the
1201 diagnostic allele.

1202 **Supplementary Table 5.** Consensus genotypes.

1203 **Supplementary Table 6.** Aneuploid, mosaic and misclassified samples.

1204 **Supplementary Table 7.** Number of samples with N and not N genotype calls in the autosomes
1205 and X chromosome of sample TL9348.

1206 **Supplementary Table 8.** Construct probe design annotation

1207

1208

1209 **Supplementary Figure 1.** Sex effect on normalized intensity for markers on chromosome X. The
1210 left figure represents 269 markers considered informative based on the lack of overlap between
1211 the distribution of intensities in males (blue) and females (red). The right represents 426
1212 makers that are not considered sex informative.

1213 **Supplementary Figure 2.** Alignments of validated construct markers. For each construct the file
1214 provides a short summary, the alignment of the working probes, the target DNA and protein
1215 sequences. The alignment of forward (black) and reverse (blue) probes is shown with the
1216 nucleotide used for “genotyping” (A) shown in red background for forward probes and in blue
1217 (T) for reverse probes. Mismatches are shown in purple background.

1218 **Supplementary Figure 3.** Examples of normal and abnormal intensity distributions. Intensity
1219 distributions for six samples with low pd_stat and six samples with high pd_stat on the
1220 autosomes and chromosome X. Colored histogram bars are the intensity values distribution on
1221 the corresponding chromosome. Colored lines are the kernel density estimates for these data.
1222 Black lines are an attempt to fit the actual data to a normal curve.

1223 **Supplementary Figure 4.** The distribution of pd_stat values in the 6,899 samples is shown on
1224 the y axis. The x axis shows the ks_stat for better contrast. Threshold determination for

1225 chromosomal sex using pd_stat. Samples in yellow were incorrectly identified as XO but are in
1226 fact XX (aka XX*, Supplementary Tables 1 and 4). Samples in green are from mouse species
1227 other than *Mus musculus*. Samples in blue are labeled Aneuploid by our algorithm. We
1228 manually established a threshold to capture all the misclassified samples and samples from
1229 other species.

1230 **Supplementary Figure 5.** Chromosome Y duplications. Spatial distribution of normalized
1231 intensity at SNPs in the proximal end of the Y chromosome in C3H/He, DBA/1 and C57BL/6
1232 samples. The range of intensities are shown in orange in cases where we had multiple samples
1233 with the duplication while samples with normal Y chromosome are shown in blue. Duplicated
1234 regions are shown in red and transition regions with uncertain copy number are shown in pink.
1235 The bottom of the figure shows the location of the MiniMUGA markers and genes.

1236 **Supplementary Figure 6.** Intensities of all construct markers present in MiniMUGA. Markers
1237 are grouped according to construct. The color denotes whether the sample is deemed to be a
1238 negative control (blue), positive control (red), or experimental (dark brown) for the respective
1239 construct. Markers with asterisks were excluded in the construct analysis.

1240 **Supplementary Figure 7.** Inbreeding thresholds. The figure shows in red the distribution of
1241 observed H calls in 385 samples representing 85 classical inbred strains. It also shows in blue
1242 the distribution of predicted number of H calls in 3,655 F1 hybrids using the consensus
1243 genotypes from 86 classical inbred strains. Tier 1 and 2 markers on the autosomes, X
1244 chromosomes and PAR were used. Thresholds for inbred, close to inbred and outbred are
1245 shown as vertical bars.

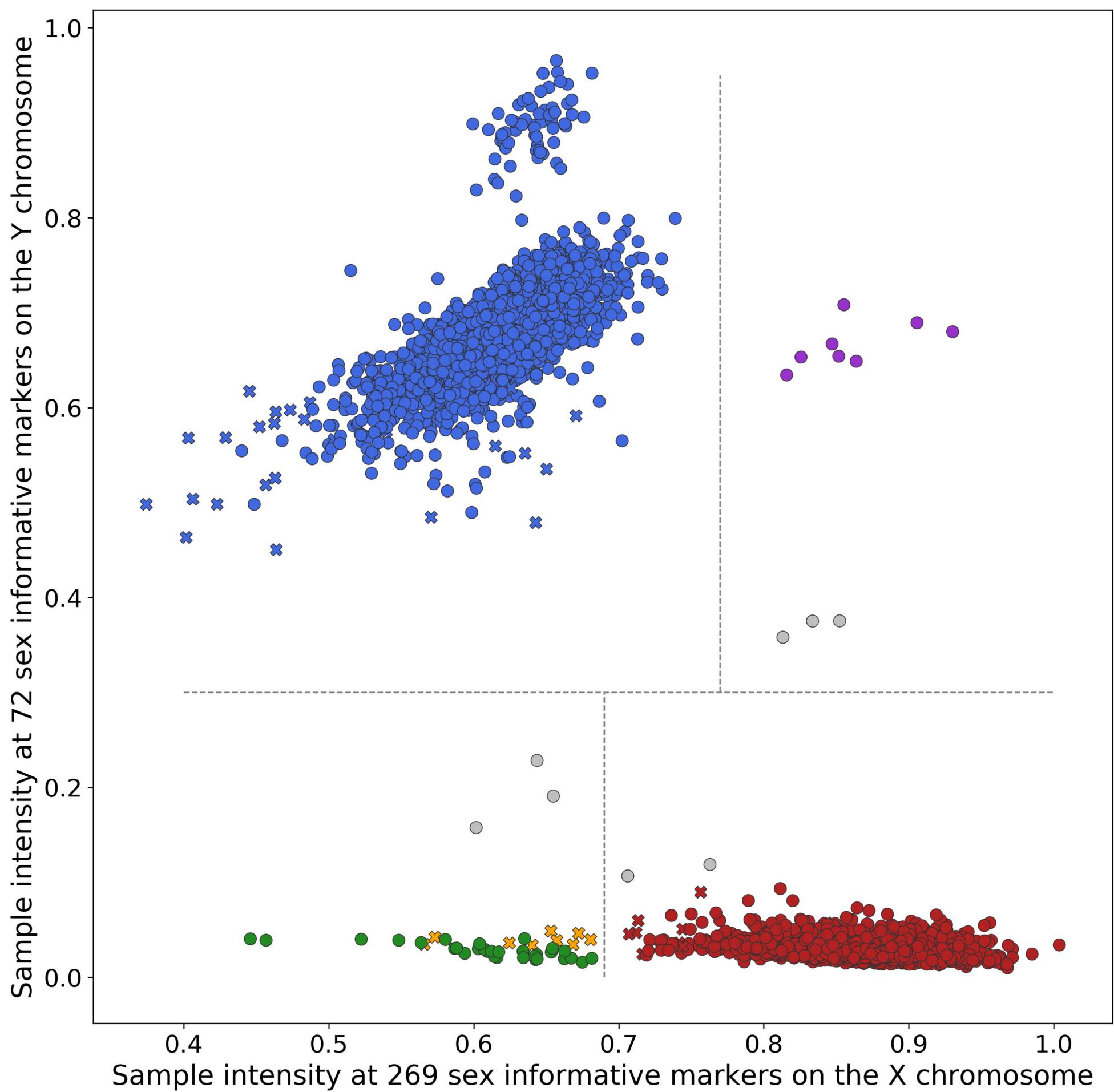
1246 **Supplementary Figure 8.** MiniMUGA Background Analysis Report for the following four female
1247 cell lines: C2Cl2, GPG C3-Tag-T1-Luc, MLE12, and C57BL/6J.

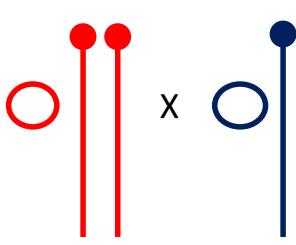
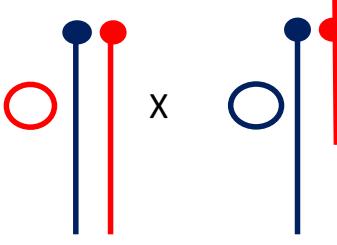
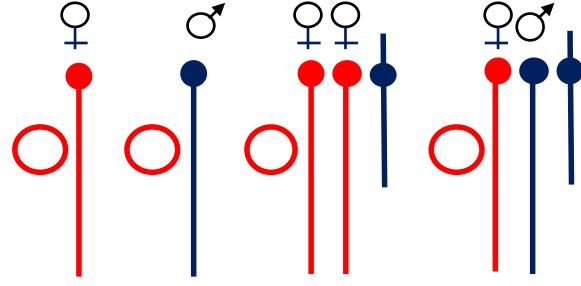
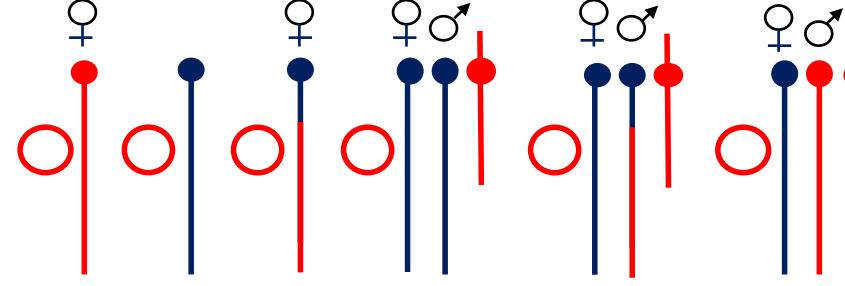
1248 **Supplementary Figure 9.** *De novo* X chromosome duplication. The range of intensities for
1249 females and males are shown in pink and blue, respectively. The sample with the duplication is
1250 shown as black line. Genotypes for the parental CC strains and the test sample are shown at
1251 the bottom as well as the first marker included in the duplication (asterisk) and the extent.

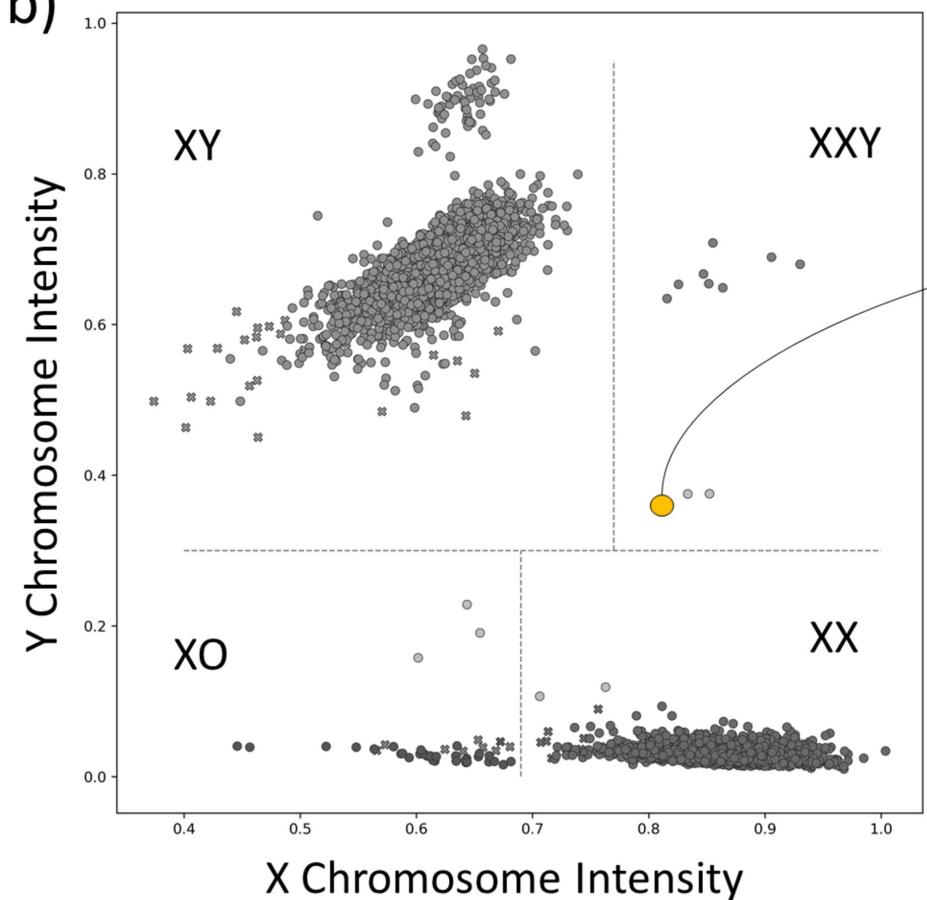
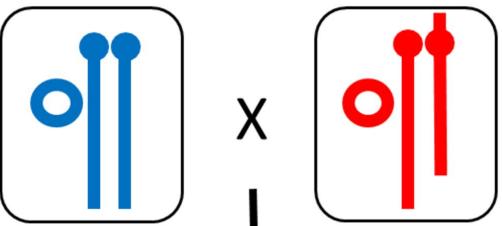
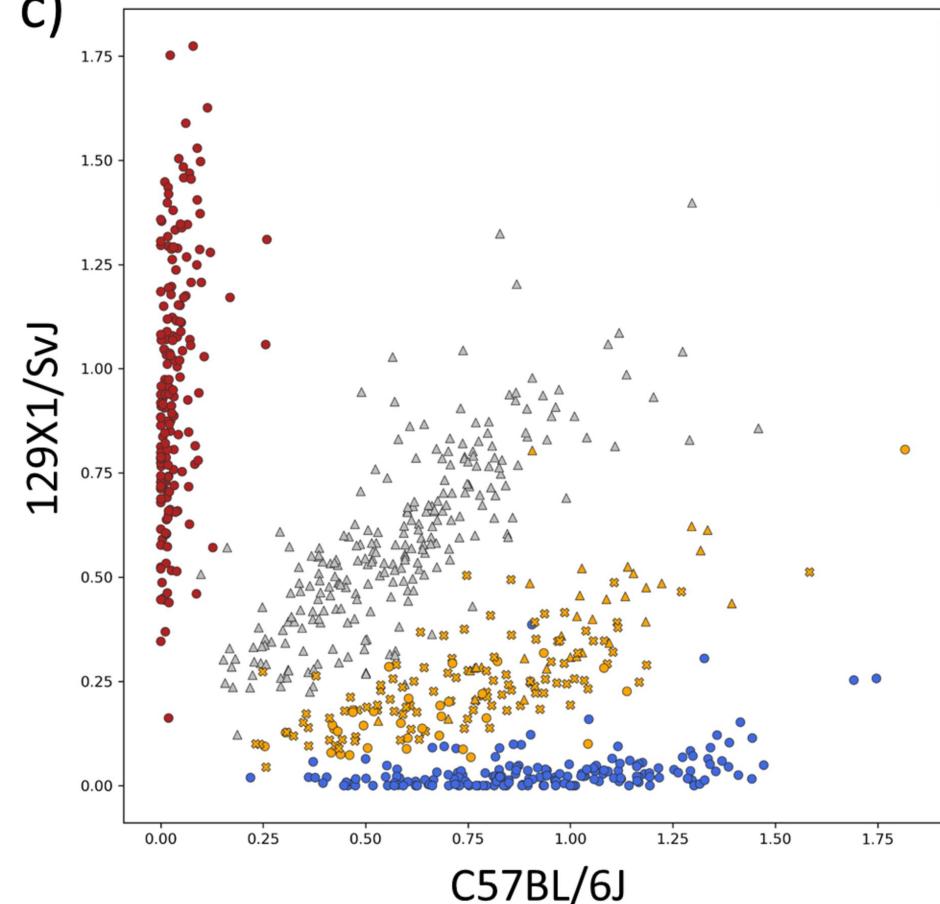
1252 **Supplementary Figure 10.** Age and breeding history of four mouse samples from the B6.129-
1253 Nox4^{tm1kkk}J congenic line maintained through breeding at UNC. Green triangles note the
1254 position of the generate allele. Red bars denote the ancestral allele for diagnostic SNPs fixed at
1255 E3. Pink bars denote ancestral alleles for diagnostic SNPs fixed at the start of the CC. Light blue
1256 bars denote diagnostic alleles at diagnostic SNPs fixed at E3. Lighter blue bars denote diagnostic
1257 alleles at diagnostic SNPs fixed at start of CC. Grey bars denote ancestral alleles at post-CC
1258 diagnostic SNPs. Dark blue bars denote diagnostic alleles at post-cc diagnostic SNPs. Split bars
1259 denote heterozygous SNPs in a sample.

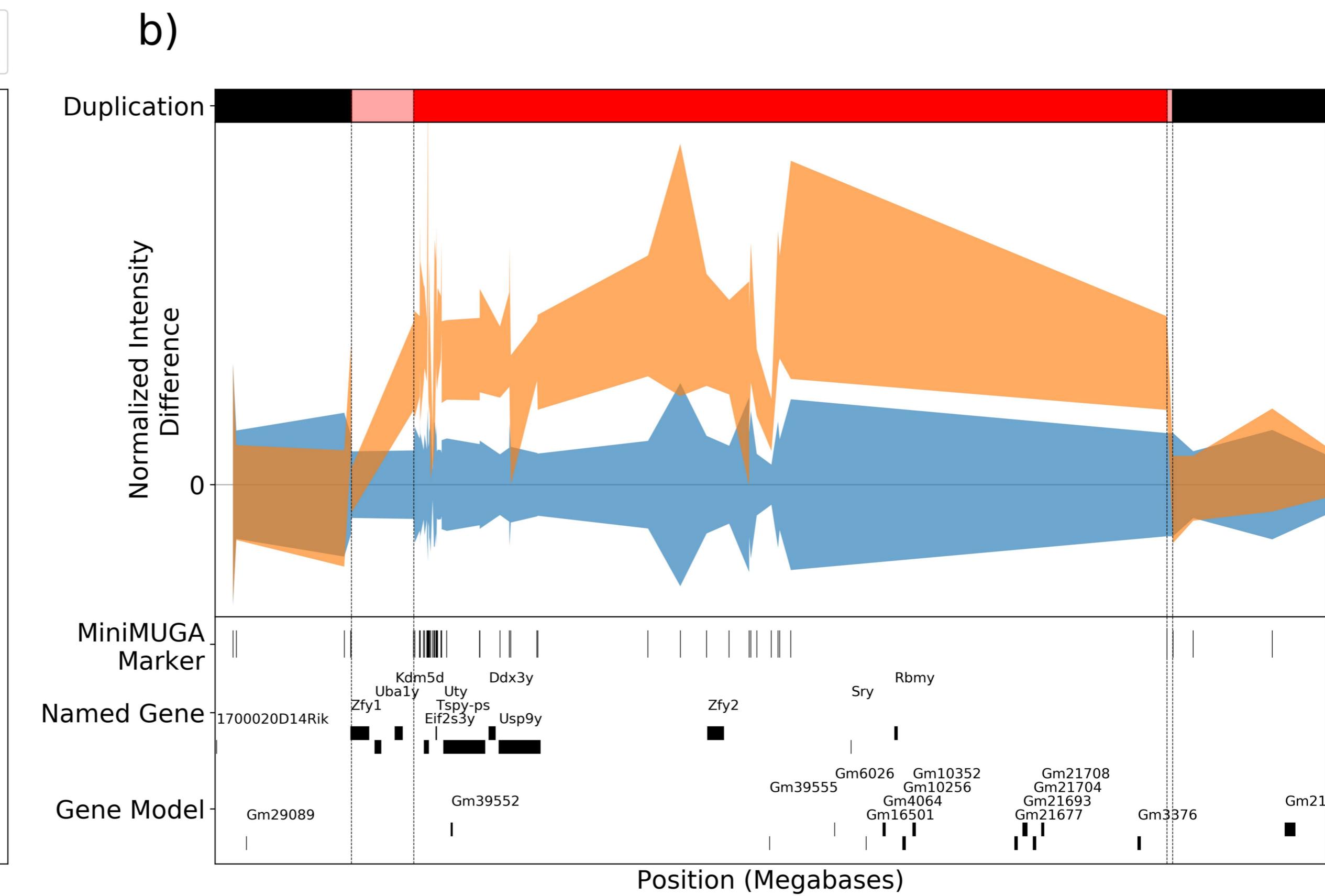
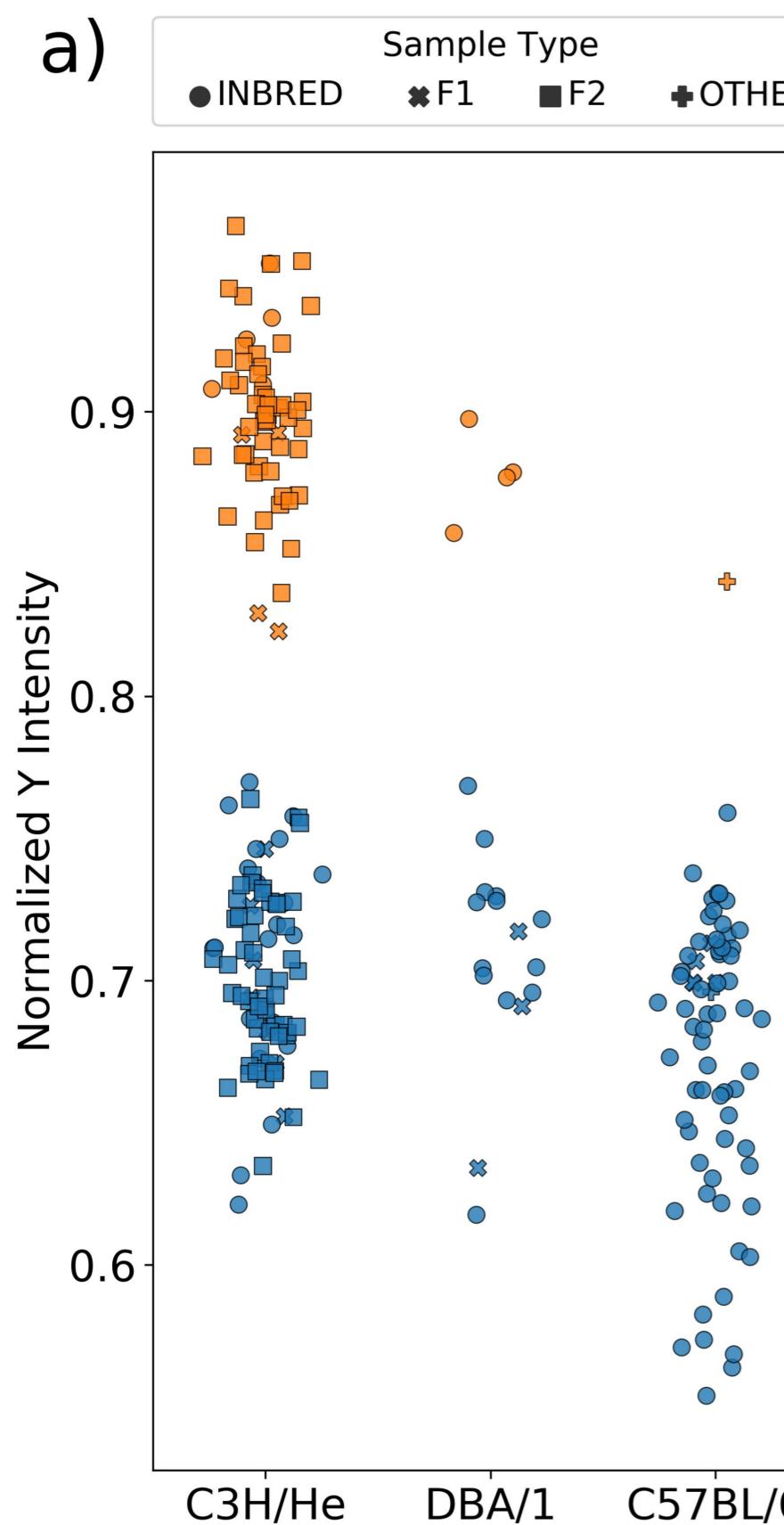
1260

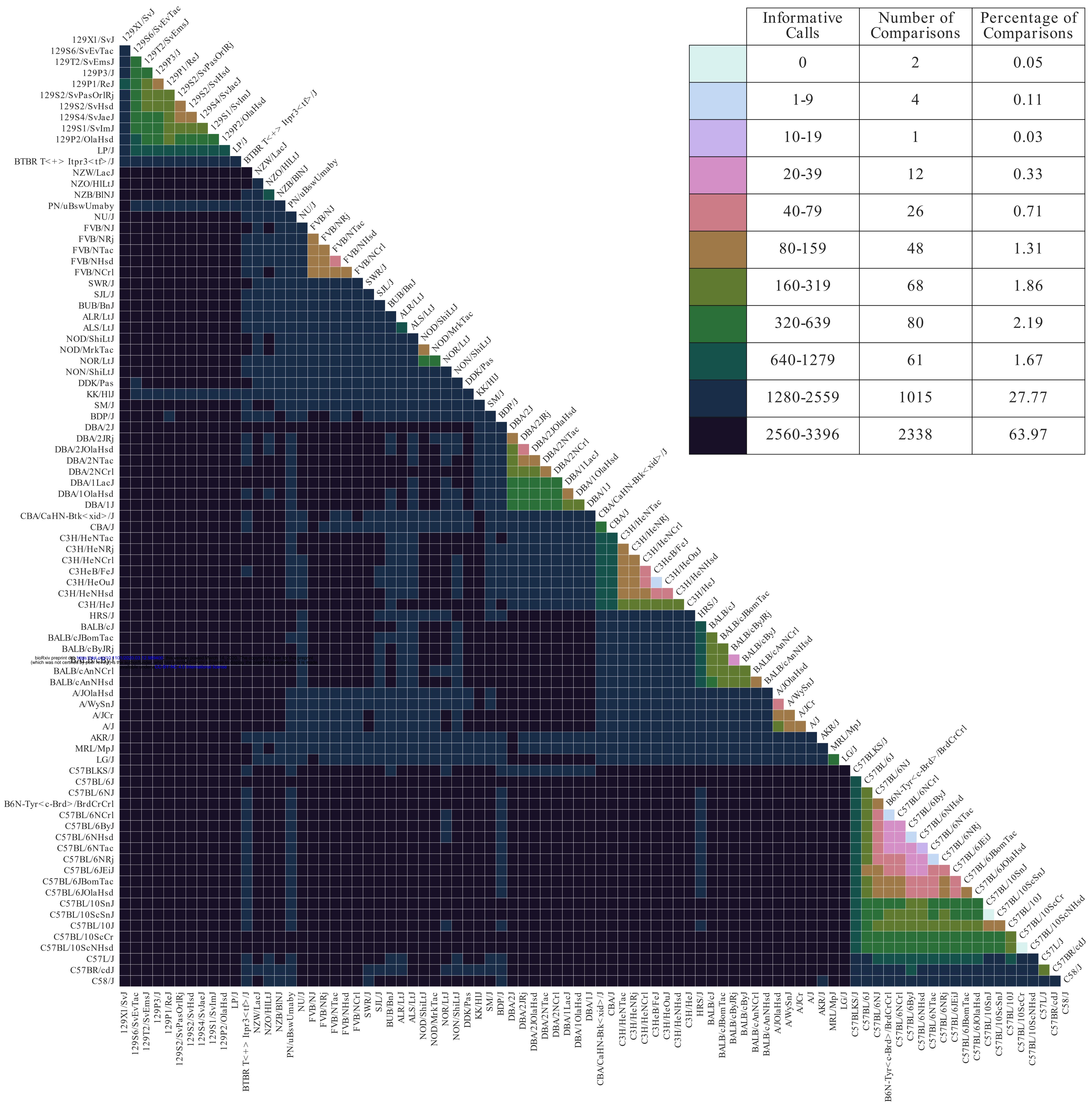
1261

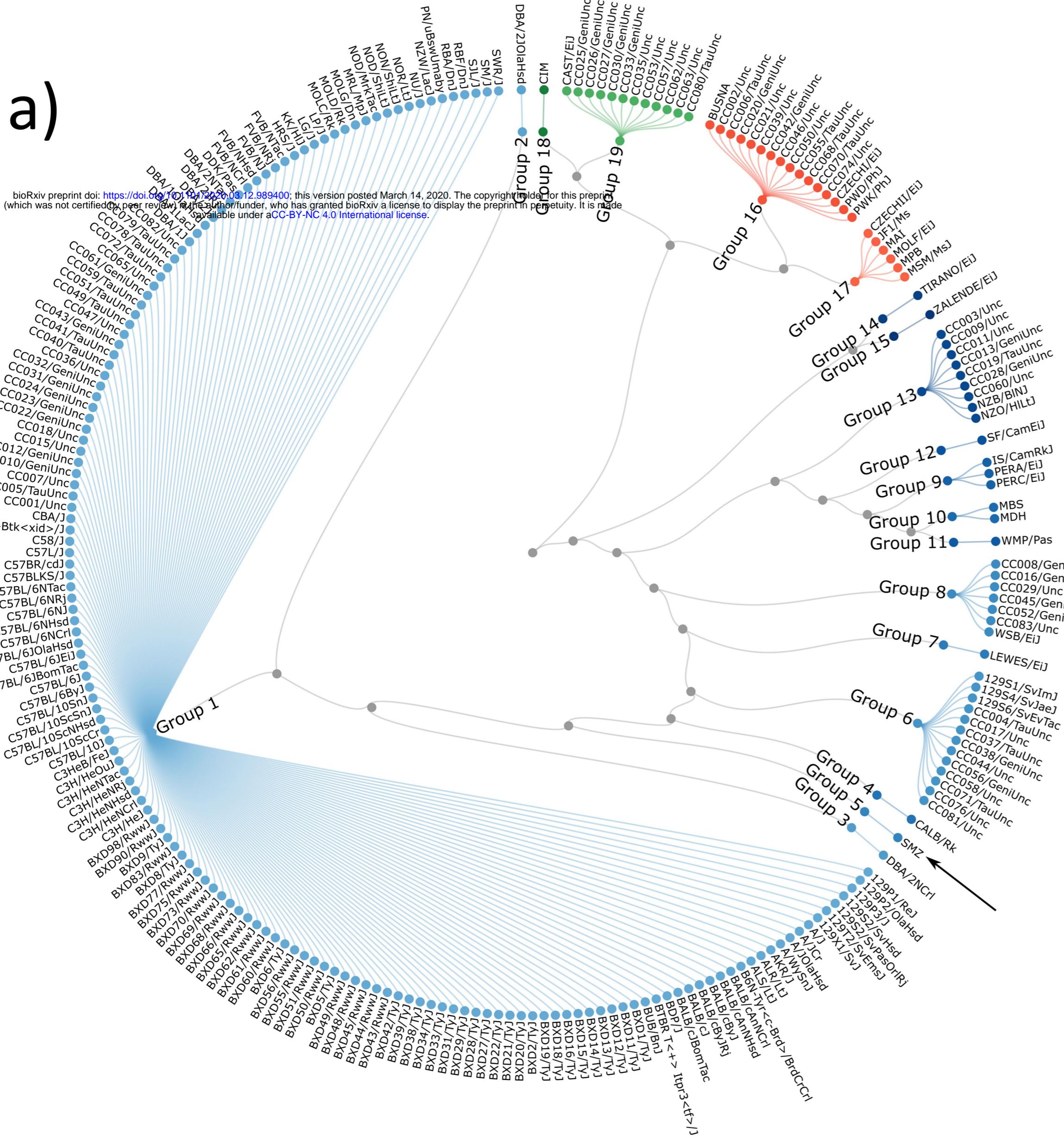
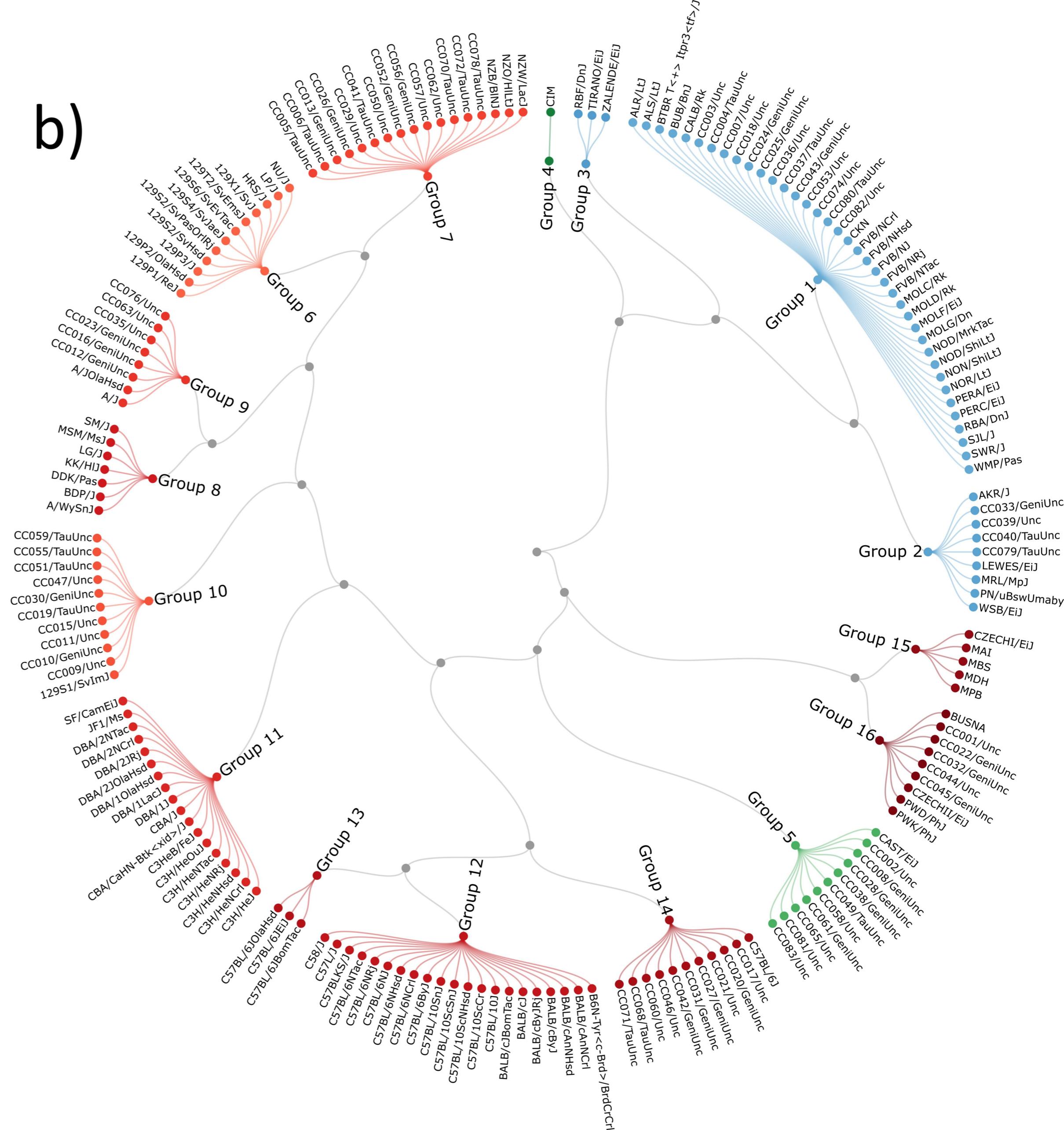


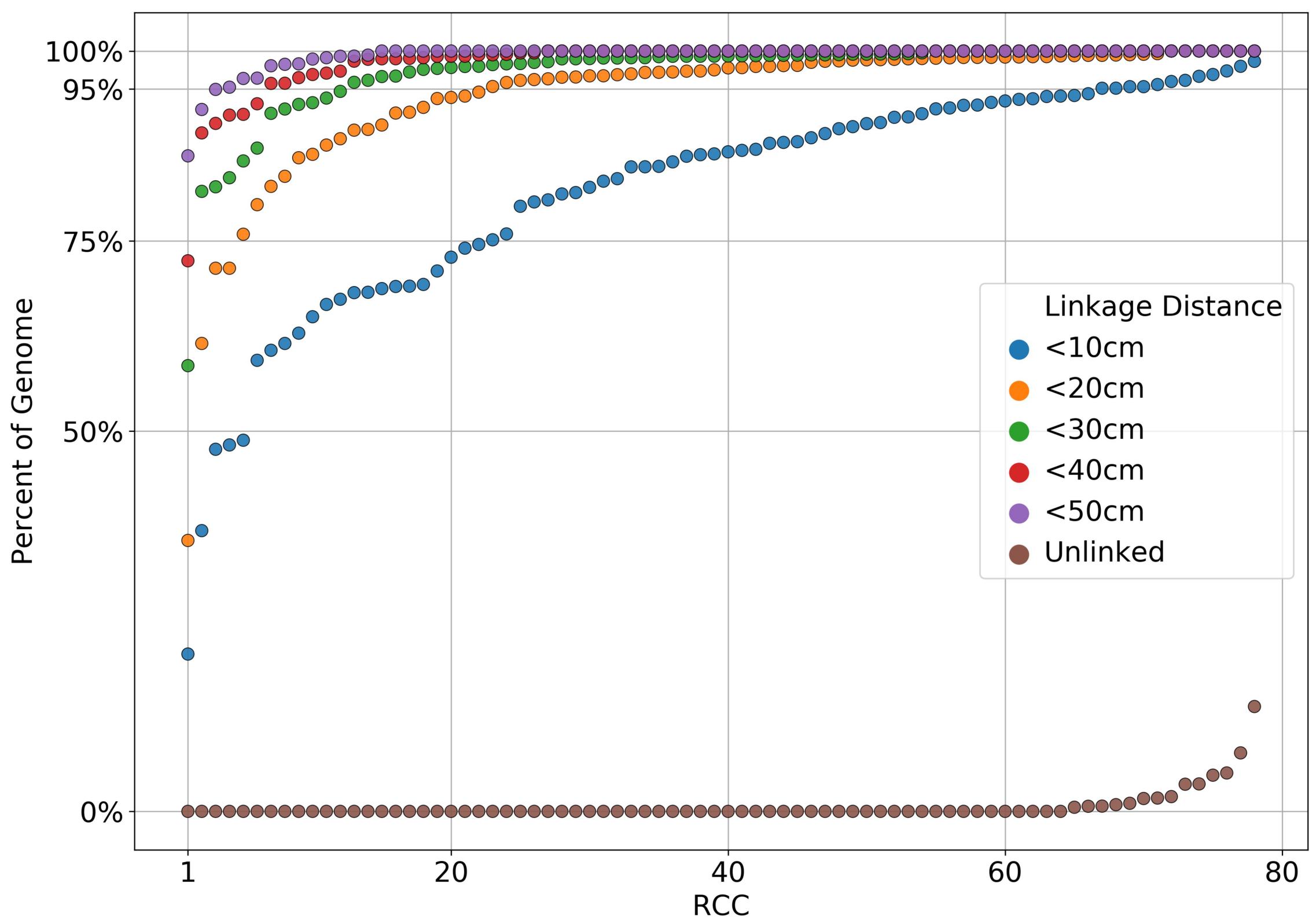
Parents		
XO and XXY Progeny		
Parental Origin	Pat Mat Mat Pat	Pat ?? Pat Pat Pat ??
Number	6 1 0 6	2 1 4 0 2 0

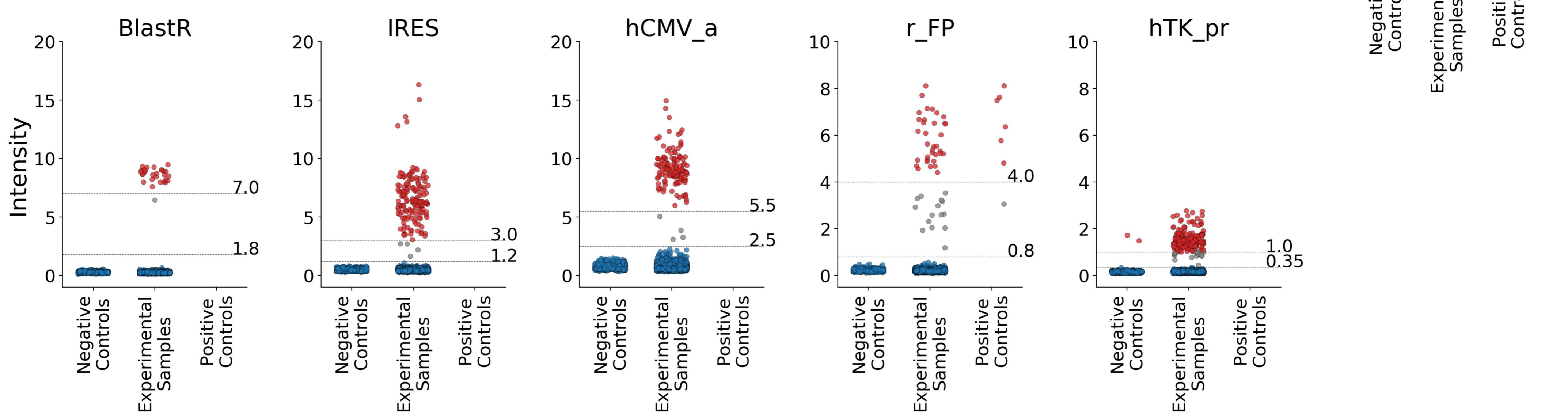
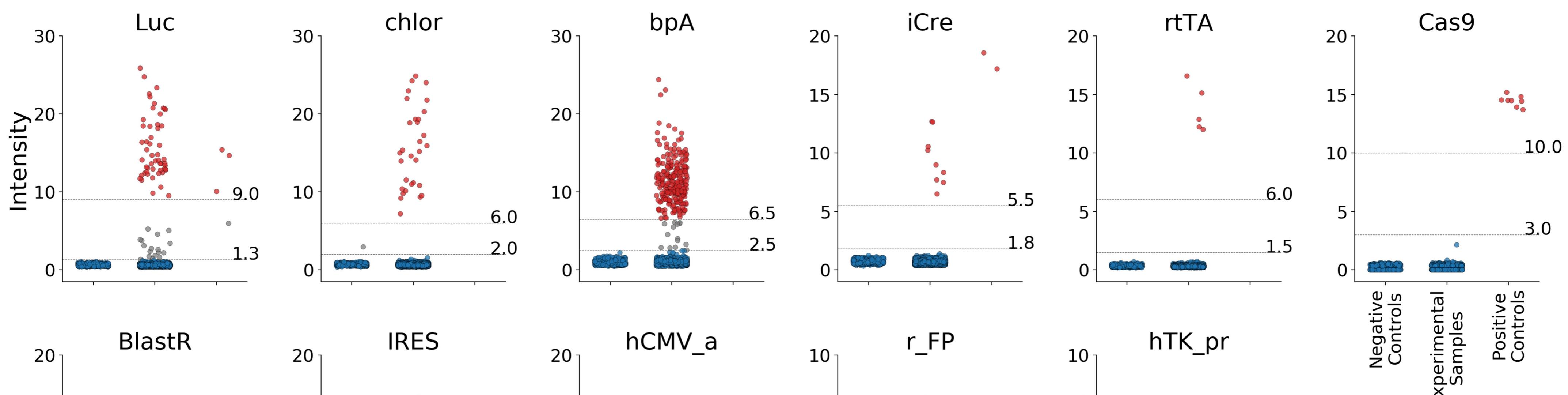
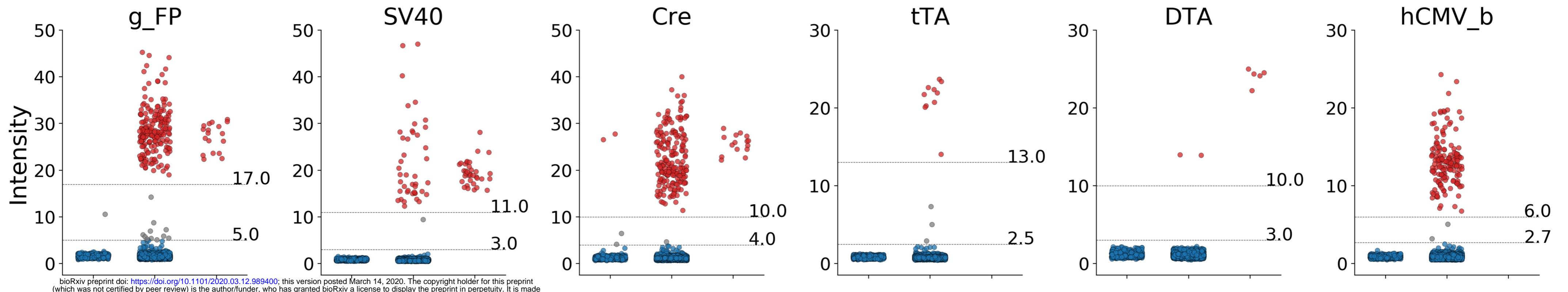
b)**a) Dam** **Sire****d)****c)**

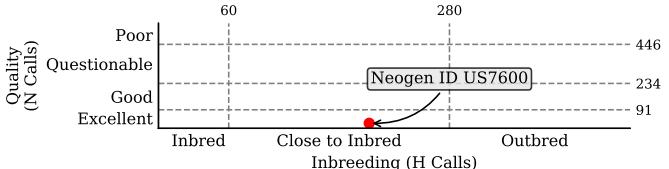
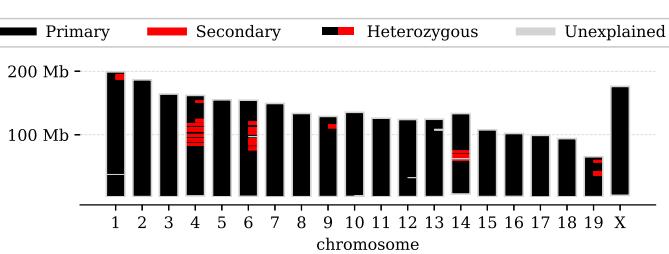


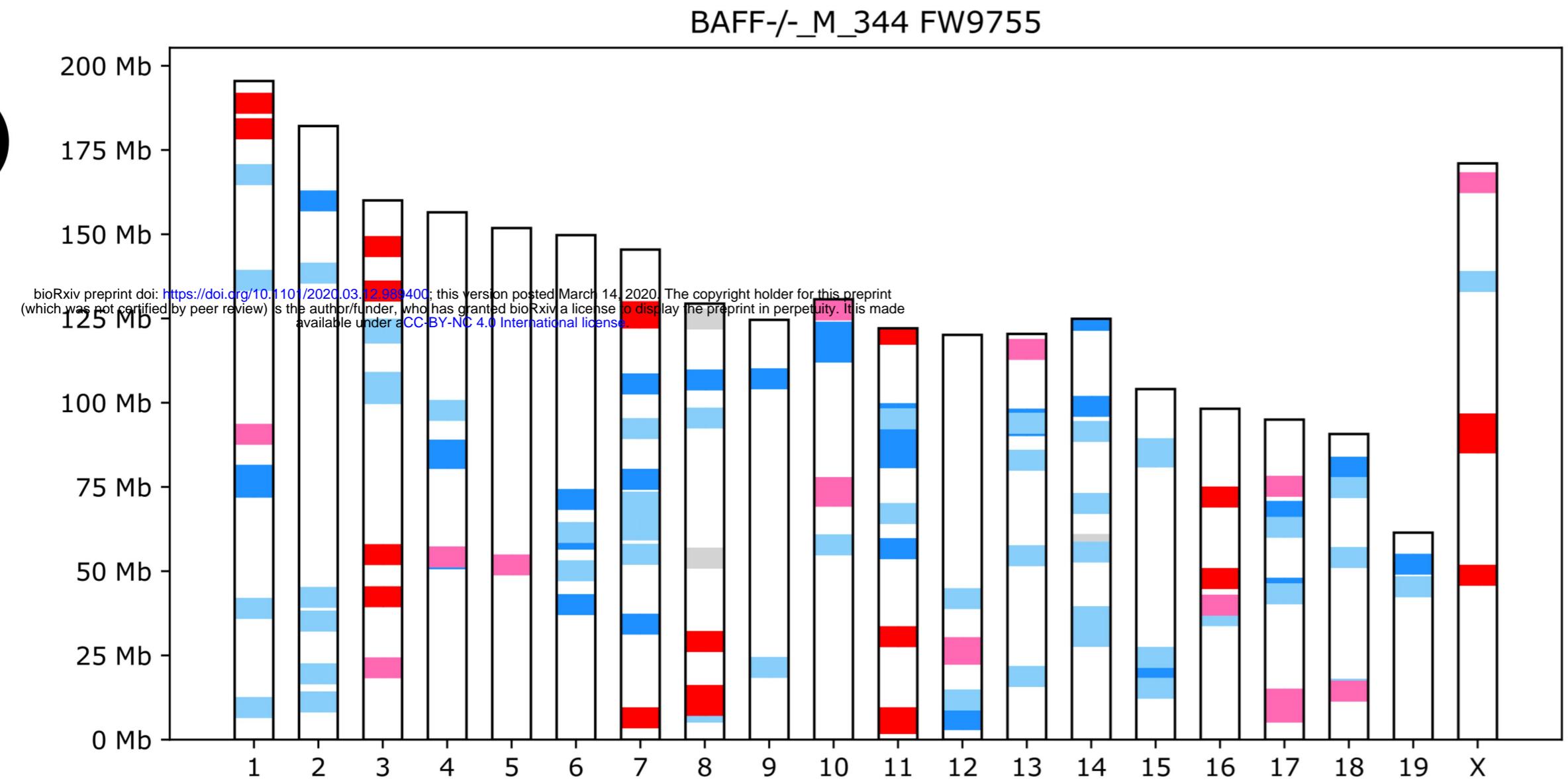


a)**b)**

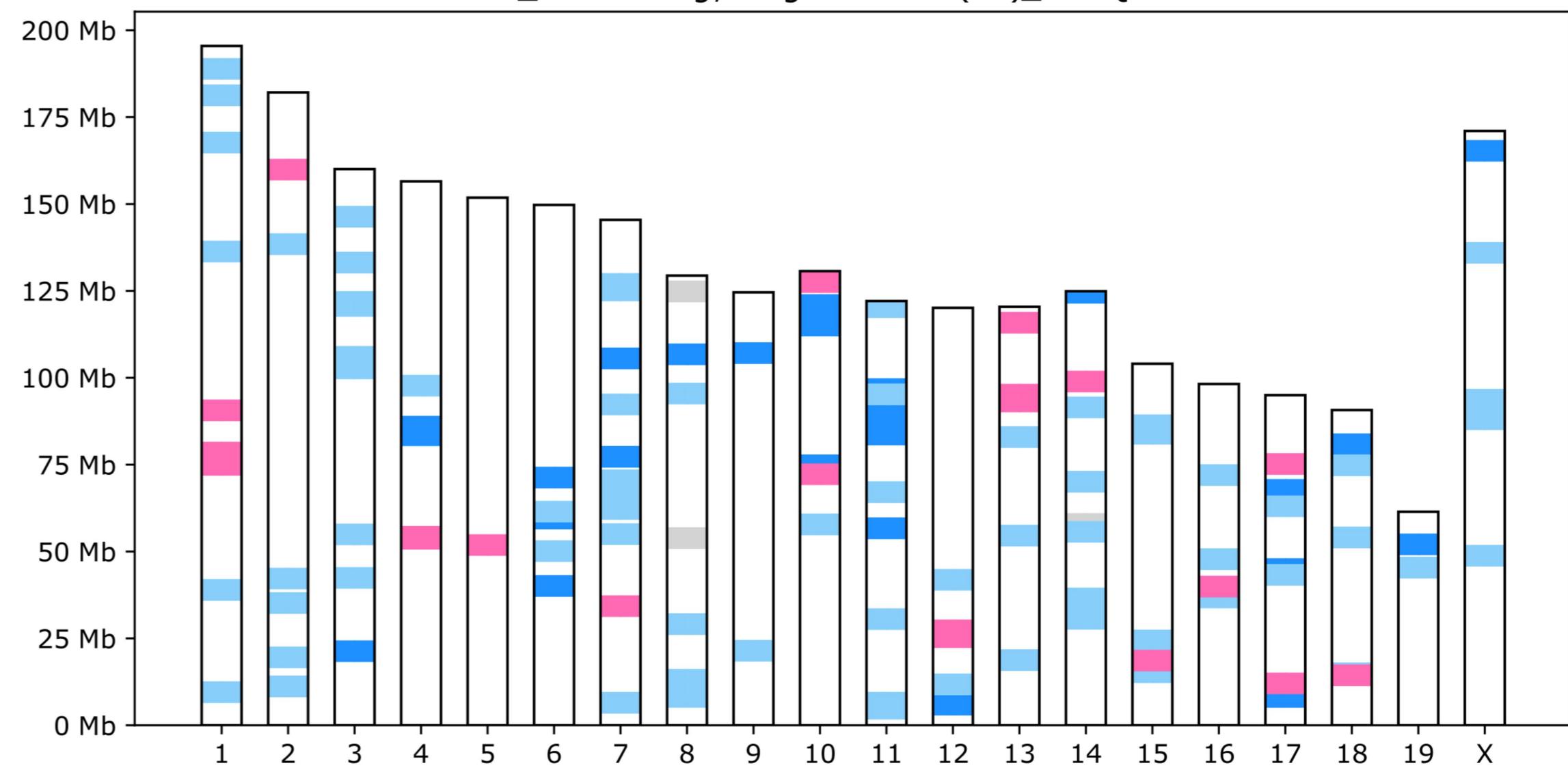
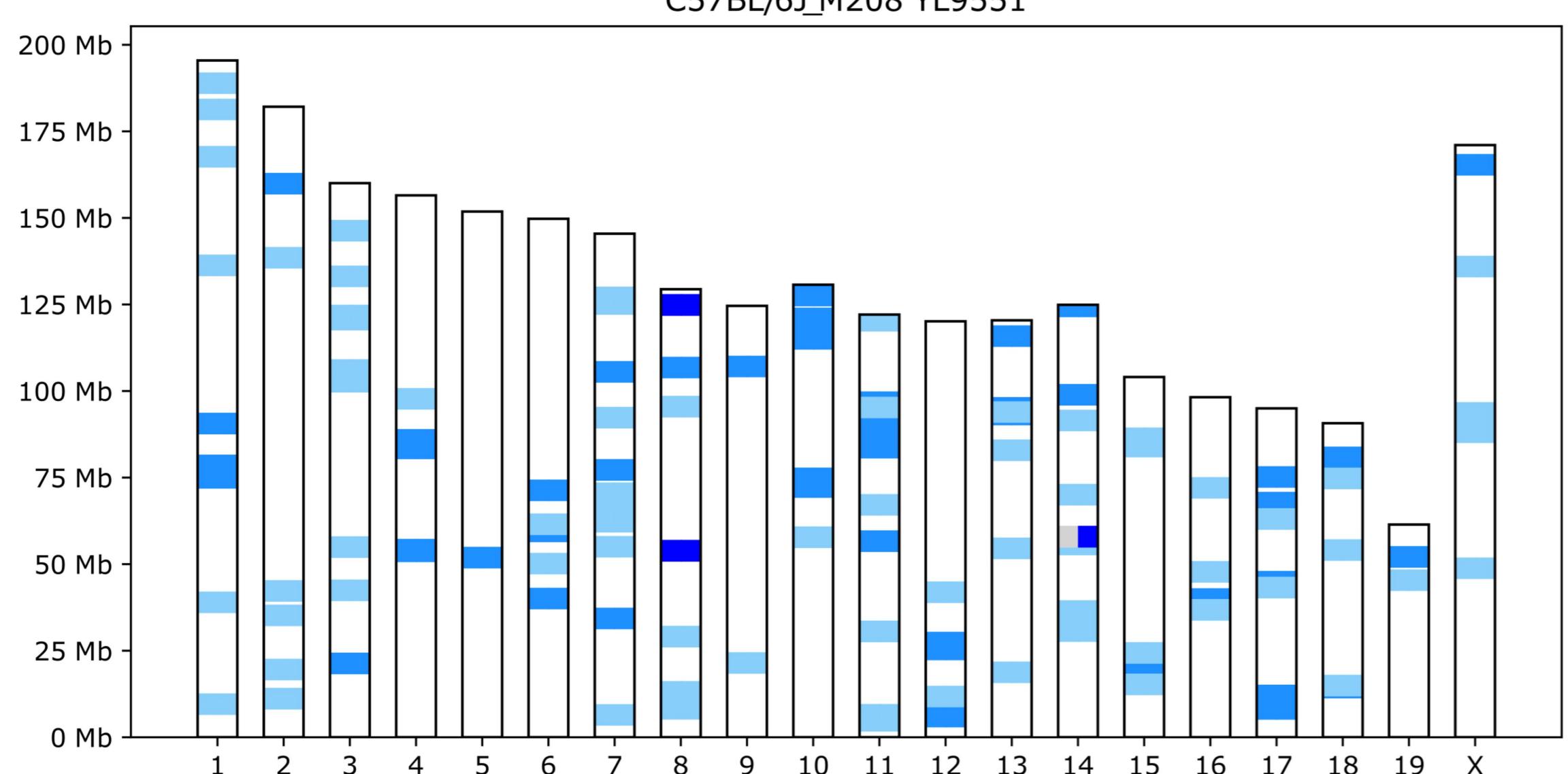




Sample ID	MMRRC_UNC_F38673																																		
Neogen ID	US7600																																		
Summary	<p>The genotype of this sample is of excellent quality. It is female and close to inbred, and likely a mix of multiple C57BL/6 substrains and (129S1/SvImJ and/or 129S2/SvHsd and/or 129S2/SvPasOrlRj and/or 129S4/SvJaeJ and/or 129S6/SvEvTac). Clustering of unexplained markers is evidence of an additional background strain.</p> <p>Diagnostic SNPs indicate the presence of the background strain groups C57BL/6 and the substrains C57BL/6J.</p> <p>The sample contains the following genetic constructs: Luciferase</p>																																		
Genotyping Quality	<p>Excellent (18 N calls) All reported results are dependent on genotyping quality.</p>																																		
Chromosomal Sex	XX																																		
Inbreeding Estimate	Close to Inbred (200 H calls at autosomal, X, and PAR chromosome markers)																																		
Inbreeding and Genotyping Quality (Plot)	 <p>Quality (N Calls)</p> <p>Inbreeding (H Calls)</p> <p>Neogen ID US7600</p>																																		
Constructs Detected	<table border="1"> <thead> <tr> <th>BlastR</th> <th>bpa</th> <th>Cas9</th> <th>chlor</th> <th>Cre</th> <th>DTA</th> <th>g_FPs</th> <th>hCMV_a</th> <th>hCMV_b</th> <th>hTK_pr</th> <th>iCre</th> <th>iRES</th> <th>Luc</th> <th>r_FPs</th> <th>rTA</th> <th>SV40</th> <th>tTA</th> </tr> </thead> <tbody> <tr> <td>-</td> <td>+</td> <td>-</td> <td>-</td> <td>-</td> </tr> </tbody> </table>	BlastR	bpa	Cas9	chlor	Cre	DTA	g_FPs	hCMV_a	hCMV_b	hTK_pr	iCre	iRES	Luc	r_FPs	rTA	SV40	tTA	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-
BlastR	bpa	Cas9	chlor	Cre	DTA	g_FPs	hCMV_a	hCMV_b	hTK_pr	iCre	iRES	Luc	r_FPs	rTA	SV40	tTA																			
-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-																			
Primary Background (Autosomes, X Chromosome)	<table border="1"> <thead> <tr> <th>Strain</th> <th>Total</th> <th>Consistent</th> <th>Inconsistent</th> <th>Heterozygous</th> <th>Excluded</th> </tr> </thead> <tbody> <tr> <td>multiple C57BL/6 substrains</td> <td>9721</td> <td>9087 (97.9%)</td> <td>50 (0.5%)</td> <td>148 (1.6%)</td> <td>436</td> </tr> </tbody> </table>	Strain	Total	Consistent	Inconsistent	Heterozygous	Excluded	multiple C57BL/6 substrains	9721	9087 (97.9%)	50 (0.5%)	148 (1.6%)	436																						
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)	<table border="1"> <thead> <tr> <th>Strain</th> <th>Total</th> <th>Explained</th> <th>Unexplained</th> <th>Excluded</th> </tr> </thead> <tbody> <tr> <td>129S1/SvImJ and/or 129S2/SvHsd and/or 129S2/SvPasOrlRj and/or 129S4/SvJaeJ and/or 129S6/SvEvTac</td> <td>198 Clustered</td> <td>182 (2.0%) Clustered</td> <td>16 (0.2%) Clustered</td> <td>0 (0.0%) Clustered</td> </tr> </tbody> </table>	Strain	Total	Explained	Unexplained	Excluded	129S1/SvImJ and/or 129S2/SvHsd and/or 129S2/SvPasOrlRj and/or 129S4/SvJaeJ and/or 129S6/SvEvTac	198 Clustered	182 (2.0%) Clustered	16 (0.2%) Clustered	0 (0.0%) Clustered																								
Strain	Total	Explained	Unexplained	Excluded																															
129S1/SvImJ and/or 129S2/SvHsd and/or 129S2/SvPasOrlRj and/or 129S4/SvJaeJ and/or 129S6/SvEvTac	198 Clustered	182 (2.0%) Clustered	16 (0.2%) Clustered	0 (0.0%) Clustered																															
Background Ideogram	 <p>200 Mb</p> <p>100 Mb</p> <p>chromosome</p> <p>Primary Secondary Heterozygous Unexplained</p>																																		
Backgrounds Detected (Diagnostic Alleles)	<table border="1"> <thead> <tr> <th colspan="5">Diagnostic Alleles Observed</th> </tr> <tr> <th>Substrain</th> <th>Homozygous</th> <th>Heterozygous</th> <th>Potential</th> <th>% Observed</th> </tr> </thead> <tbody> <tr> <td>C57BL/6J</td> <td>77</td> <td>45</td> <td>156</td> <td>78.2%</td> </tr> </tbody> </table> <table border="1"> <thead> <tr> <th colspan="5">Strain Group</th> </tr> <tr> <th>Strain Group</th> <th>Homozygous</th> <th>Heterozygous</th> <th>Potential</th> <th>% Observed</th> </tr> </thead> <tbody> <tr> <td>C57BL/6 (B6N-Tyr/BrdCrCrl, C57BL/6J, C57BL/6BomTac, C57BL/6JEiJ, C57BL/6OlaHsd, C57BL/6NCrl, C57BL/6NHsd, C57BL/6NJ, C57BL/6NRj, C57BL/6NTac)</td> <td>6</td> <td>1</td> <td>21</td> <td>33.3%</td> </tr> </tbody> </table>	Diagnostic Alleles Observed					Substrain	Homozygous	Heterozygous	Potential	% Observed	C57BL/6J	77	45	156	78.2%	Strain Group					Strain Group	Homozygous	Heterozygous	Potential	% Observed	C57BL/6 (B6N-Tyr/BrdCrCrl, C57BL/6J, C57BL/6BomTac, C57BL/6JEiJ, C57BL/6OlaHsd, C57BL/6NCrl, C57BL/6NHsd, C57BL/6NJ, C57BL/6NRj, C57BL/6NTac)	6	1	21	33.3%				
Diagnostic Alleles Observed																																			
Substrain	Homozygous	Heterozygous	Potential	% Observed																															
C57BL/6J	77	45	156	78.2%																															
Strain Group																																			
Strain Group	Homozygous	Heterozygous	Potential	% Observed																															
C57BL/6 (B6N-Tyr/BrdCrCrl, C57BL/6J, C57BL/6BomTac, C57BL/6JEiJ, C57BL/6OlaHsd, C57BL/6NCrl, C57BL/6NHsd, C57BL/6NJ, C57BL/6NRj, C57BL/6NTac)	6	1	21	33.3%																															

a)

JW_SMARTA-Tg; IFNgR1<tm1>(B6)_F23 QV8571

b)**c)**

Content	Chromosomal set	Inbred	F1	CC	Cross	Unclassified	Cell lines	Total
Initial	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
Final	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

Background	Strain Group	Diagnostic Type	Full	Partial
129P2/OlaHsd	129P	substrain	25	0
129P3/J	129P	substrain	54	0
129S1/SvImJ	129S	substrain	82	13
129S2/SvHsd	129S	substrain	7	1
129S2/SvPasOrlRj	129S	substrain	36	0
129S4/SvJaeJ	129S	substrain	45	0
129S5/SvEvBrd	129S	substrain	12	0
129S6/SvEvTac	129S	substrain	41	0
129T2/SvEmsJ	129T	substrain	38	0
129X1/SvJ	129X	substrain	39	0
A/J	A	substrain	58	7
A/JCr	A	substrain	53	0
A/JOlaHsd	A	substrain	38	0
BALB/cAnNCrl	BALB /c	substrain	36	2
BALB/cAnNHsd	BALB /c	substrain	109	4
BALB/cByJ	BALB /c	substrain	3	4
BALB/cByJRj	BALB /c	substrain	19	0
BALB/cJ	BALB /c	substrain	103	3
BALB/cJBomTac	BALB /c	substrain	47	0
C3H/HeJ	C3H/He	substrain	166	2
C3H/HeNCrl	C3H/He	substrain	39	0
C3H/HeNHsd	C3H/He	substrain	39	1
C3H/HeNRj	C3H/He	substrain	42	0
C3H/HeNTac	C3H/He	substrain	45	14
C57BL/6J	C57BL/6	substrain	136	20
C57BL/6JBomTac	C57BL/6	substrain	41	2
C57BL/6JOlaHsd	C57BL/6	substrain	43	0
C57BL/6NJ	C57BL/6	substrain	37	7
C57BL/6NRj	C57BL/6	substrain	20	0
B6N-Tyr<c-Brd>/BrdCrCrl	C57BL/6	substrain	21	10
DBA/1J	DBA/1	substrain	70	0
DBA/1LacJ	DBA/1	substrain	77	2
DBA/1OlaHsd	DBA/2	substrain	32	0
DBA/2J	DBA/2	substrain	112	0
DBA/2JOlaHsd	DBA/2	substrain	39	0
DBA/2JRj	DBA/2	substrain	30	0
DBA/2NCrl	DBA/2	substrain	85	14
DBA/2NTac	DBA/2	substrain	36	10
FVB/NCrl	FVB	substrain	47	0
FVB/NHsd	FVB	substrain	39	1
FVB/NJ	FVB	substrain	72	7
FVB/NRj	FVB	substrain	47	0

FVB/NTac	FVB	substrain	37	0
NOD/MrkTac	NOD	substrain	33	0
NOD/ShiLtJ	NOD	substrain	51	3
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
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
Subtotal			635	0

TOTAL	2916	127
--------------	-------------	------------

WGS
Sanger
UNC
Sanger
UNC
UNC
UNC
UNC
Sanger
UNC
UNC
UNC
Sanger
UNC
Sanger
UNC
Sanger
UNC
UNC
UNC
UNC
Reference
UNC
UNC
Sanger
UNC
UNC
Sanger
UNC
UNC
Sanger
UNC
UNC
Sanger
UNC
UNC
UNC
Sanger
UNC
UNC
UNC
UNC
Sanger
UNC

UNC
UNC
Sanger

Abraham
Sanger

Name	Abbreviation
"Greenish" Fluorescent Protein (EGFP, EYFP, ECFP)	g_FP
SV40 large T antigen	SV40
Cre recombinase	Cre
Tetracycline repressor protein	tTA
Diphtheria toxin	DTA
Human CMV enhancer <i>version b</i>	hCMV_b
Luciferase and firefly luciferase	Luc
Chloramphenicol acetyltransferase	chloR
Bovine growth hormone poly A signal sequence	bpA
iCre recombinase	iCre
Reverse improved tetracycline-controlled transactivator	rtTA
Caspase 9	cas9
Blasticidin resistance	BlastR
Internal Ribosome Entry Site	IRES
hCMV enhancer <i>version a</i>	hCMV_a
"Reddish" fluorescent protein (tdTomato, mCherry)	r_FP
Herpesvirus TK promoter	hTK_pr

Total

# of probes	# of distinct probes
19	19
18	18
16	12
14	14
11	11
10	7
10	10
9	9
8	4
8	8
8	4
7	7
6	4
6	6
5	4
6	6
2	2

163	145
------------	------------

A)	C57BL/6J		DBA/2J		C57BL/6 group	DBA/2J group	C57BL/6	DBA/2	Other
	Full	Partial	Full	Partial					
I	0	0	4	0	2	24	1	0	0
II	72	0	68	0	4 (2)	0	0	0	1*
III	34	0	16	0	0	0	0	0	0
IV	30	20	24	0	0	0	0	0	0

B)	A/J		C57BL/6J		129S1/SvImJ		NOD/ShiLtJ	
	Full	Partial	Full	Partial	Full	Partial	Full	Partial
PreCC	47	0	116	7	75	6	34	0
During CC	8	3	16	7	2	4	2	0
PostCC	0	2	0	3	0	1	0	1