PROTEIN CODING VARIATION IN OUTBRED LABORATORY MOUSE STOCKS PROVIDES A MOLECULAR BASIS FOR DISTINCT RESEARCH APPLICATIONS.

Belinda Cornes[1], Carolyn Paisie[1], Emily Swanzey[1], Andrew Schile[1], Kelly Brackett[1], Laura Reinholdt*[1], Anuj Srivastava*[1]

1. The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609

*co-corresponding
laura.reinholdt@jax.org
anuj.srivastava@jax.org

keywords: outbred stocks, laboratory mice, reproducibility, exome, genetic quality control

running title: Exome sequencing of outbred laboratory mouse stocks

ABSTRACT

Outbred laboratory mice (*Mus musculus*) are readily available and have high fecundity, making them a popular choice for biomedical research, especially toxicological and pharmacological applications. Direct high throughput genome sequencing (HTS) of these widely used research animals is an important genetic quality control measure that ensures research reproducibility. HTS data have been used to confirm the common origin of outbred stocks and to molecularly define distinct outbred populations. But these data have also revealed unexpected population structure and homozygosity in some populations; genetic features that emerge when outbred stocks are not properly maintained. We used exome sequencing to discover and interrogate protein coding variation in a newly established population of Swiss-derived outbred stock (J:ARC) that is closely related to other, commonly used CD-1 outbred populations. We used these data to describe the genetic architecture of the J:ARC population including heterozygosity, minor allele frequency, LD decay, and we defined the novel, protein-coding sequence variation. These data reveal the expected genetic architecture for a properly maintained outbred stock and provide a basis for on-going genetic quality control. We also compared these data to protein coding variation found in a multiparent outbred stock, the Diversity Outbred (J:DO). We found that the more recently derived, multiparent outbred stock has significantly higher interindividual variability, greater overall genetic variation, higher heterozygosity, and fewer novel variants than the Swiss derived J:ARC stock. However, among the novel variants found in the J:DO stock, significantly more are predicted to be protein-damaging. That individuals from this population can tolerate a higher load of potentially damaging variants highlights the buffering effects of allelic diversity and the differing selective pressures in these stocks. While both outbred stocks offer significant individual heterozygosity, our data provide a molecular basis for their intended applications, where the J:DO are best suited for studies requiring maximum, population-level genetic diversity and power for mapping, while the J:ARC are best suited as a general-purpose outbred stock with robust fecundity, relatively low allelic diversity, and less potential for extreme phenotypic variability.

INTRODUCTION

There are 1,374 inbred strains and outbred laboratory mouse stocks cited in the biomedical research literature (BULT *et al.* 2019). Genome sequencing of inbred strains has confirmed the expected homozygosity and intrastrain genetic divergence that are the results of decades of inbreeding and reproductive isolation (MOUSE GENOME SEQUENCING *et al.* 2002; KEANE *et al.* 2011; YALCIN *et al.* 2011; CHANG *et al.* 2017; SRIVASTAVA *et al.* 2017; LILUE *et al.* 2018). Because inbred mice are genetically identical, interindividual phenotypic variability can be attributed to experimental variables and to a lesser extent, epigenetic variation (WEICHMAN AND CHAILLET 1997). These unique features make inbred mouse strains a popular choice for biomedical researchers who endeavor to minimize genetic variability and to quantify experimental (extrinsic, non-genetic) variability. However, inbred strains do not recapitulate the extent of interindividual variation found in a human cohort, patient group, or population making them inappropriate for studies that seek to model such variation.

In contrast to inbred strains, outbred stocks exhibit interindividual variation and heterozygosity, coupled with higher fecundity. The four most widely cited outbred stocks are CF-1, Swiss Webster, NMRI, and CD-1 (BULT *et al.* 2019), but all trace their origins to a colony at The Rockefeller Institute that was established with 2 male and 7 female 'Swiss' mice imported from the Pasteur Institute in 1926 (CHIA *et al.* 2005). Over the ensuing decades, mice from these stocks were imported by commercial breeders, NIH, and academic institutions where colonies were ideally maintained with at least 25 breeding pairs to maintain a coefficient of inbreeding (F) of less than 1% per generation with selection for maximum fecundity (FESTING *et al.* 1972) (CHIA *et al.* 2005). A variety of approaches have been used to examine the genetic variation resident in these outbred stocks including alloenzyme analysis and other biochemical approaches (GROEN AND LAGERWERF 1979; RICE AND O'BRIEN 1980; CUI *et al.* 1993), high-density SNP panels (ALDINGER *et al.* 2009; YALCIN *et al.* 2010), and most recently low coverage genome sequencing (YALCIN *et al.* 2010; NICOD *et al.* 2016). Data from these studies provide estimates of minor allele frequency, heterozygosity, and coefficients of inbreeding. Specifically, low coverage sequencing of these stocks has revealed 1) relatively low allelic diversity given their common origin, 2) appreciable levels of heterozygosity of common variants (0.20-0.35) in properly maintained populations, and 3) relatively few novel variants (<5%) relative to the

C57BL/6J reference genome and other common inbred laboratory strains (YALCIN *et al.* 2010; NICOD *et al.* 2016).

The JAX Swiss Outbred stock was imported to The Jackson Laboratory (JAX) from The Animal Resources Centre (ARC) in Canning Vale, Western Australia in 2020 as a group of 128 individuals (64 male and 64 female). To fully characterize the extent of protein coding variation in this new outbred population, we sequenced the exomes of the 32 male and 32 female G3 animals that were used to initiate new breeding funnels at JAX (J:ARC). We used these data to estimate inbreeding, heterozygosity, and interindividual variation, and to molecularly define the colony. These metrics confirmed the outbred nature of this new colony of CD-1 stock, and additionally reveal 5,105 (3.8%) novel protein coding and splice site variants, which falls in the range of the proportion of novel variants found in other outbred colonies and stocks (YALCIN *et al.* 2010; NICOD *et al.* 2016). To provide a molecular genetic quality control standard for outbred stocks, we compared the J:ARC exome call sets to exome variant call sets from very different outbred stock, derived from a multiparent outbred population, the Diversity Outbred (J:DO). We found that the J:DO have higher population level genetic variation (~3.6X), higher interindividual variation, higher heterozygosity, and higher allelic diversity due to their multi-substrain parental origins. We found fewer novel variants in the J:DO, yet these novel variants are predicted to have higher impact on protein function. These common and distinct genomic features of the J:ARC and J:DO can serve as the basis for estimating sample size, where fewer J:ARC mice are needed to capture the full range of segregating variation in the population, but a similarly sized cohort of J:DO mice offers more interindividual genetic variation and higher population level genetic variation. Both types of outbred stocks can be used for genetic mapping, however, their distinct genetic architectures must be accounted for in study design as previously shown (GATTI *et al.* 2014; NICOD *et al.* 2016), and should be regularly monitored to ensure reproducibility.

MATERIALS & METHODS

Samples

J:ARC(S), JAX Swiss Outbred. The JAX Swiss Outbred stock was imported to The Jackson Laboratory from The Animal Resources Centre (ARC), Canning Vale, Western Australia in 2020.

123 The imported mice (G0) were paired into 64 breeding units, and sperm and eggs were
124 harvested from the G1 offspring. To establish each of the 32 breeding funnels for live colony
125 maintenance in the JAX vivarium barrier, two distinct units were selected for reciprocal *in vitro*
126 fertilization and IVF-generated embryos were pooled. The resulting live-born animals from each
127 funnel were designated G2 (Figure 1). These mice were subsequently bred according to the
128 Poiley rotational breeding scheme to produce 32 breeder pair units / funnels at G3(POILEY
129 1960). Spleen samples were collected the 64 G3 J:ARC (J:ARC(S), RRID:IMSR_JAX:034608,
130 JAX stock number 034608) (32 females and 32 males) that populated each unit (Supplementary
131 Data 1).
132
133 *J:DO, JAX Diversity Outbred.* We collected spleen from 20 J:DO, Diversity Outbred
134 (RRID:IMSR_JAX:009376, JAX stock #9376) mice from The Jackson Laboratory, 10 female and
135 10 male, non-sibling mice, breeding generation 43 (Supplementary Data 1).

136 ## DNA Isolation, Exome Library Preparation, and Sequencing

137 DNA isolation, exome library preparation, and sequencing were performed by Genome
138 Technologies at The Jackson Laboratory. DNA was isolated from spleen using the NucleoMag
139 Tissue Kit (Machery-Nagel) according to the manufacturer's protocol. DNA concentration and
140 quality were assessed using the Nanodrop 8000 spectrophotometer (Thermo Scientific), the
141 Qubit Flex dsDNA BR Assay (Thermo Scientific), and the Genomic DNA ScreenTape Analysis
142 Assay (Agilent Technologies). Mouse exome libraries were constructed using the KAPA
143 HyperPrep Kit (Roche Sequencing and Life Science) and SureSelect XT Community Design
144 Mouse All Exon v2 Target Enrichment System (Agilent Technologies), according to the
145 manufactures' protocols. Briefly, the protocol entails shearing the DNA using the E220
146 Focused-ultrasonicator (Covaris), size selection targeting 400 bp, ligating Illumina specific
147 adapters, and PCR amplification. The amplified DNA libraries are then hybridized to the Mouse
148 All Exon probes (Agilent Technologies) and amplified using indexed primers. The quality and
149 concentration of the libraries were assessed using the High Sensitivity D1000 ScreenTape
150 (Agilent Technologies) and KAPA Library Quantification Kit (Roche Sequencing and Life
151 Science), respectively, according to the manufacturers' instructions. Libraries were sequenced
152 150 bp paired-end on an Illumina NovaSeq 6000 using the S4 Reagent Kit v1.5.

Sequence Data Analysis

An overview of the sequence data analysis pipeline is shown in Supplementary Figure 1. All reads were subjected to quality control using an in-house QC script. Samples with base qualities greater ≥ 30 over 70 % of read length were used in the downstream analysis. High-quality reads were mapped to the mouse genome (build-mm10) using BWA-mem (bwa-0.7.9a) at default parameters (ALDINGER *et al.* 2009). The resulting alignment was sorted by coordinates and converted to binary alignment map (BAM) format by Picard v 2.8.1- SortSam utility (http://picard.sourceforge.net). Picard-MarkDuplicates module was used to remove the duplicates from data. The Genome Analysis Toolkit (GATK v4) (MCKENNA *et al.* 2010; DEPRISTO *et al.* 2011) module BaseRecalibrator was used to pre-process the alignments. Target capture efficiency was determined using Picard-HsMetrics (1.95). The recalibrated bam alignment file was used to input GATK-Haplotype Caller at parameters -stand_call_conf 50.0, -stand_emit_conf 30.0 and variants calls were restricted to the target region (Mouse All Exon v2). Finally, raw variants calls were soft filtered using GATK VariantFiltration (DP < 25, QD < 1.5 and FS > 60), annotated by snpEff 3.6.c (CINGOLANI *et al.* 2012) and the highest impact variant was reported by GATK VariantAnnotator. All variants were further annotated with mouse dbSNP v150.

GATK-HaplotypeCaller in GVCF mode was used for joint genotyping of J;DO, J:ARC, and combined (J:DO and J:ARC) samples. The CombineGVCF utility was used to gather all the samples and then executed the GenotypeGVCF command. All analyses were performed on jointly called variants and only included variants found in both “J:ARC only” and “J:DO only” called datasets (N: 120,864). This effectively revealed the differences between the two datasets and directly provided the data needed to resolve genotypes.

Heterozygosity

To observe the distribution of mean heterozygosity across all samples, Plink was used (PURCELL *et al.* 2007). The `--het` option in Plink was used to create heterozygous information for each sample which included the observed number of homozygous genotypes `[O(Hom)]` and the number of non-missing genotypes `[N(NM)]`. This information was then used to calculate the observed heterozygosity rate per individual using the formula `(N(NM) - O(Hom))/N(NM)`.

185 To calculate per variant heterozygosity for variants shared between J:DO and J:ARC, the
186 number of samples in which a shared variant was heterozygous was divided by the total
187 number of genotyped samples. The inbreeding coefficient (F_H), was calculated
188 using *VCFtools* (–het) as previously described (STOFFEL *et al.* 2016; FOSTER *et al.* 2021).
189 PLINK was used to calculate r^2 for all pairs of autosomal SNPs called from joint genotyping in
190 the J:ARC samples (425,409) and the J:DO samples (117,429). SNPs that were missing in
191 more than 5% of the samples and that were monomorphic were removed. The 95th percentile of
192 r^2 values for SNPs spaced up to 1 Mb apart were selected.

193 Allele Frequency

194 To observe differences between allele frequencies across the genome between J:DO and
195 J:ARC samples, the minor allele frequency (MAF) was calculated separately for J:DO and
196 J:ARC samples at each variant site using the `--freq` option in Plink.

197 Clustering/Dendrogram

198 Principal Component Analysis (PCA) plots were performed in Plink. PCA is a multivariate
199 statistical method used to produce any number of uncorrelated variables (or principal
200 components) from a data matrix containing observations across a number of potentially
201 correlated variables. The principal components are calculated so that the first principal
202 component accounts for as much variation in the data as possible in a single axis of variation
203 (component), followed by additional components. Here the variants called in both J:ARC and
204 J:DO datasets were used to observe the similarities within and between datasets.
205
206 To show similarity groups among the J:ARC samples, a dendrogram was performed using the
207 ggdendro package in R-3.6.2 (https://cran.r-project.org). A dendrogram is a tree diagram
208 showing hierarchical clustering that represents the relationships of similarity among a group of
209 entities.
210

211 Novel Variants

212 The joint genotyping variant call file (vcf) of J:ARC and J:DO were flagged for known variants in
213 dbSNP150 (SHERRY *et al.* 2001), European Variation Archive (EVA) (CEZARD *et al.* 2022) , Sanger

214 mouses genome project (KEANE *et al.* 2011) (ftp://ftp-mouse.sanger.ac.uk//REL-2004-v7-
215 SNPs_Indels/mgp_REL2005_snps_indels.vcf.gz), mouse Collaborative Cross genome
216 (ftp://ftp.sra.ebi.ac.uk/vol1/ERZ460/ERZ460702/Merged_69_flagged.tab.vcf.gz) (SRIVASTAVA *et*
217 *al.* 2017), and Diversity Outbred low pass sequencing (in-house resource) using vcftools 'vcf-
218 annotate' (KEANE *et al.* 2011) and snpsift_4.2 (CINGOLANI *et al.* 2012). We also flagged J:ARC
219 variants for their presence in J:DO joint and single sample variant call sets and did a similar
220 analysis for J:DO variants. The variants not present in any of these resources were considered
221 novel and were further annotated by snpeff v4.3 (CINGOLANI *et al.* 2012) using the mouse
222 GRCm38.75 snpeff database. Finally, highest effect variants are selected by gatk-3.6
223 VariantAnnotator (MCKENNA *et al.* 2010). Functional gene and pathway annotations (including
224 KEGG, GO, UP, BIOCARTA, REACTOME) for the genes harboring novel high impact variants
225 were compiled using DAVID Bioinformatics tools. DAVID Bioinformatics tools were then used
226 to identify clusters of genes with related functional annotations using a medium classification
227 stringency and a Benjamini score of >0.05 as the significance threshold for enrichment
228 (SHERMAN *et al.* 2022)

229 Results and Discussion

230 The J:ARC outbred stock currently maintained at The Jackson Laboratory was
231 established through the importation of mice from the Australian Animal Resource Center (ARC)
232 in 2020. The origins of this stock are CD-1 mice that were acquired by ARC from Charles River
233 Laboratories in 1991 and 2005. Like most outbred laboratory mice, CD-1 mice trace their
234 origins to 2 male and 7 female "Swiss" mice that were obtained by The Rockefeller Institute in
235 1926 (CHIA *et al.* 2005). The live J:ARC colony was initiated with founder animals (G3) in 32
236 breeding funnels. We sought to determine the level of inbreeding and relatedness in the
237 founder animals of this colony and to catalog the coding variants segregating in each funnel for
238 future use in genetic quality control monitoring (STROBEL *et al.* 2015). Since some outbred
239 stocks differ quite dramatically, we sought to compare the genetic architecture of the outbred
240 J:ARC to a multiparent outbred population (J:DO) to demonstrate the differing genetic
241 architectures of these stocks and to establish a reference set of genetic variants for genetic
242 monitoring. The J:DO mice are derived from eight founder inbred strains representing the three
243 major subspecies of laboratory mice, *M.m.musculus, M.m.domestics*, and *M.m.castaneous*.
244 This population captures 90% of the genetic variation in laboratory mice and with each

245 generation of breeding, sufficient accumulated recombination to ensure high LD decay, and
246 thereby maximum power for genetic mapping (GATTI *et al.* 2014; SAUL *et al.* 2019). This
247 population is maintained at the Jackson Laboratory through 175 breeding funnels using
248 MateSel(KINGHORN AND KINGHORN 2021) to minimize inadvertent phenotypic selection, to
249 optimize diversity, and to prevent intracolony structure in each generation based on pedigreed
250 population history.

251 *Genetic variation.* We generated exome data from 64 J:ARC mice (32 male and 32 female)
252 representing 32 breeding funnels at G3 and from 20 J:DO mice (10 male and 10 females) at
253 generation 43. We generated ~203M high-quality reads on average, and after alignment to the
254 reference genome, the mean target coverage was 162X (91% target covered at 30X) for J:ARC
255 samples and 155X (90% target covered at 30X) for J:DO samples. Overall, 90% of the target
256 exome regions were covered by 30 or more reads in both sets of samples (see Supplementary
257 Figure 2).
258
259 Through joint genotyping, we identified 478,011 and 135,825 variants in J:DO and
260 J:ARC samples, respectively, where a variant is defined as a SNP/INDEL present in a
261 sequenced sample when compared to the C57BL/6J, inbred mouse reference genome (mm10,
262 GRCm38). To adjust for the difference in sample size between the J:DO and J:ARC datasets
263 and any potential influence this could have on overall discovery rates, we randomly subset the
264 J:ARC samples to a set of twenty from joint genotyping results and then recorded the total
265 number of variants present in the subset. We repeated this analysis over 100 iterations and
266 found from this bootstrapping method that the J:DO harbor, on average, 3.6X more variants
267 overall than J:ARC. This higher allelic diversity is expected in the J:DO since the stock is a
268 multiparent population derived from eight diverse founder inbred strains.
269
270 *Heterozygosity and inbreeding.* To evaluate heterozygosity in these two outbred populations,
271 we estimated sample level heterozygosity for the entire set of variant calls (Figure 2A). The
272 average, individual heterozygosity in J:ARC is 24%, which is consistent with previously
273 estimates from low pass WGS for CFW outbred populations (NICOD *et al.* 2016) (YALCIN *et al.*
274 2010). In contrast, sample level heterozygosity in the J:DO samples was notably higher (nearly
275 40%) than J:ARC. Given that previous analyses have estimated J:DO haplotype-level
276 heterozygosity at around 80% (https://www.jax.org/-/media/jaxweb/files/jax-mice-and-

277 services/009376_J_DO_Mar2021.pdf), it is important to note that we used reference-based
278 genotype calls and we calculated heterozygosity across the entire call set. Previously,
279 heterozygosity for J:DO mice was calculated on a limited set of variants based on genotype
280 probabilities of eight parental genotypes called from known polymorphic SNPs from the Mouse
281 Universal Genotyping Array (MUGA) series (MORGAN *et al.* 2015; SIGMON *et al.* 2020). To
282 recapitulate this haplotype-based approach using our exome variant calls, we used 3007 calls
283 at markers from GigaMUGA genotyping array (MORGAN *et al.* 2015) and for this subset of
284 variants, we found heterozygosity of ~82% consistent with previous array-based estimates
285 (see Supplementary Figure 3).
286
287 To compare the variant level heterozygosity between the two populations, we identified
288 the variants that are common between J:ARC and J:DO and then for each variant determined
289 the number of samples with heterozygous genotypes (Figure 2B). We found that a lower
290 proportion of J:ARC samples are heterozygous for any given variant than the J:DO, and that
291 there is greater intraindividual variability in the proportion of heterozygous variants in the J:DO
292 population. To further explore these differences in the distribution of heterozygosity, we
293 calculated the minor allele frequencies (MAF) for each of the common variants across the
294 genome (Figure 3A-C). We found that for each common variant, the MAF in the J:ARC was
295 consistently lower (mean 0.18, median 0.13) than the J:DO (mean 0.28, median 0.28) and
296 similar to previously published MAF for variants segregating in properly maintained outbred
297 CD-1 mouse populations (YALCIN *et al.* 2010).
298
299 Variant data can also be used to estimate the coefficient of inbreeding (F), which is a
300 metric that describes the distribution of variants in a population. According to the Hardy-
301 Weinberg (HW) principle, F=0 for a given variant indicates that it is in HW equilibrium. By this
302 measure, a fully inbred strain has an inbreeding coefficient of F=1.0 (100%). When calculated
303 across all J:ARC variants and samples, the average level of inbreeding is -.033 (-3.3%), which
304 falls within expectations for an outbred population (YALCIN *et al.* 2010). Though fewer samples
305 were used to generate the J:DO dataset, the inbreeding coefficient calculated from the J:DO
306 was closer to HW equilibrium at -0.007 (-0.7%). These F coefficients indicate that both
307 populations are maximally outbred. The negative values are the result of heterozygous calls
308 that exceed HW expectations which in these datasets could be due to erroneous genotype
309 calls or low sampling of the populations. While the coefficients of inbreeding for these

310 populations are low, there are regions of the genome where shared variants are fixed. These
311 regions of homozygosity are distributed through the J:ARC and J:DO genomes (Figure 4). It is
312 likely that some homozygous variant calls (regions of homozygosity, ROH) are the result of
313 heterozygous deletions. Because our methodology doesn't distinguish between potential
314 hemizygous and true homozygous variant calls, we may be overestimating the extent of ROH
315 in these genomes.
316
317 *Interindividual variability*. An important consideration in selecting an outbred laboratory mouse
318 strain and sample size in experimental design is interindividual genetic variability. Taking
319 advantage of our variant data, we used principal component analysis to assess interindividual
320 and interstrain variation in the J:ARC and J:DO. Using all jointly called variants, we found the
321 expected interstrain variation (PC1) but more interindividual variability in J:DO samples than
322 J:ARC (PC2) (Figure 5A). These data show that greater genetic diversity can be achieved with
323 fewer samples in the J:DO by at least an order of magnitude. Unbiased clustering of the J:ARC
324 variant data recapitulated the known pedigree and sibling relationships of the J:ARC samples
325 (Figure 5B), highlighting the utility of variant data for pedigree reconstruction in outbred
326 populations.
327
328 *LD Decay.* Linkage disequilibrium (LD) describes the degree to which variants co-segregate in
329 a population. Meiotic recombination leads to new haplotypes and is the molecular mechanism
330 that drives LD decay. Pairwise analysis of linked markers in an interval is used to assess LD
331 decay in a population and these data are useful for inferring swept radius for marker spacing in
332 genetic association studies. Higher LD decay in a mapping population provides higher
333 resolution for genetic mapping. We examined this facet of genetic architecture in the J:ARC
334 and J:DO populations and while our sample size was low, we found that both populations offer
335 LD decay sufficient for genetic mapping, with higher LD decay in the J:DO ($LD_{1/2}$ = ~0.1 Mb)
336 (Figure 5C). Our estimate of LD decay is higher than previous estimates for J:DO, but this is
337 consistent with the more advanced breeding generation used in our study (GATTI *et al.* 2014).
338 LD decay in the J:ARC is consistent with previous estimates for commonly used laboratory
339 outbred populations ($LD_{1/2}$ = ~0.3 Mb) *(YALCIN et al. 2010).*
340
341 *Novel Variants.* To identify the novel variant calls in the J:ARC and J:DO, we compared the
342 calls to all published variants from all sequenced laboratory mouse strains in dbSNP, EVA, the

343 Sanger mouse genomes project (LILUE *et al.* 2018), Mouse Collaborative Cross Genome
344 (SRIVASTAVA *et al.* 2017), and to in house J:DO genome variation call sets (unpublished). Of the
345 135,825 and 478,011 variants found in J:ARC and J:DO respectively, 96.24 and 99.08% are
346 variants that have been identified in other laboratory mouse strains. We identified 5,105 [2,191
347 > AF 0.2] and 4,308 [1,888 > AF 0.2] novel variants respectively (Table 1). The larger number of
348 novel variants in J:ARC mice is attributable to the relative lack of published variant data from
349 Swiss derived outbred stocks compared to the complete catalog of variation available from the
350 eight founder inbred strains of the Collaborative Cross recombinant inbred panel from which
351 the J:DO were derived. Functional annotation of the novel variants revealed that J:DO have
352 nearly twice the number of protein-damaging variants (including exon deletions, frameshift,
353 stop gain / loss, splice acceptor / donor sites, start loss) ((J:DO:797 [329 > AF 0.2, HIGH],
354 J:ARC: 408 [151 > AF 0.2, HIGH]) (Figure 6). The details of each variant are provided in
355 Supplementary Data 2. The genes harboring these high impact variants in J:DO show
356 functional enrichment for post-translational modifications (Ubl conjugation, isopepeptide
357 bond), DNA repair, and translational regulation / RNA binding; while the genes harboring high
358 impact variants in J:ARC do not show functional enrichment. The novel protein damaging
359 variants in J:DO are variants that have naturally arisen in the population but have not been lost
360 through purifying selection or through artificial selection. These genetic differences are
361 attributable not only to the distinct origins and ages of these outbred stocks, but to the
362 different breeding paradigms and selective pressures that have occurred during their
363 maintenance. For example, many outbred mouse stocks and populations have been subject to
364 artificial selection for desirable laboratory traits (fecundity, docility, size, etc.) leading to lack of
365 variability for some phenotypes. One example is body weight where variation is less desirable
366 because it complicates pharmacological applications, i.e. dosing studies. As a result, body
367 weight variation in CD-1 and related outbred like J:ARC is minimized, however this is not the
368 case in J:DO where a larger range of body weights is observed (Supplementary Figure 4).

369 CONCLUSIONS

370 We used exome sequencing to define protein coding variation in two outbred populations of
371 mice with distinct origins. We found that the multiparent, J:DO outbred population has more
372 than 3X population-level variation than the CD-1 derived J:ARC outbred population. Both
373 populations have coefficients of inbreeding that are consistent with previous estimates,

374 confirming that both are maximally outbred. Our analysis of variants that are shared between
375 the J:ARC and J:DO shows that individual heterozygosity in both populations is high, but allelic
376 diversity, intraindividual variability, and minor allele frequency are higher in the J:DO
377 population. For research applications this means that similarly sized cohorts of J:DO mice will
378 provide more genetic diversity than J:ARC, making the former population the better choice for
379 studies that endeavor to maximize genetic diversity. Our data and multiple previous studies
380 show that both outbred populations are useful for genetic mapping. But the higher genome-
381 wide LD decay in J:DO will confer higher mapping resolution.

382 Overall, we found more novel variants in J:ARC compared to J:DO and this is likely due
383 to the relative paucity of published / accessible variant data for commonly used outbred
384 strains, especially CD-1 from which the J:ARC population was derived. While there were fewer
385 novel variants in the J:DO, more of these were variants that are predicted to be protein
386 damaging and potentially more likely to reduce survival and/or fecundity. While this study
387 doesn't allow us to make any inferences about the load of deleterious alleles and their
388 respective contributions to fecundity, litter size is a trait that is frequently used to estimate
389 fecundity and breeding data from The Jackson Laboratory show that J:DO females have litter
390 sizes that are comparable to many common laboratory inbred strains (mean litter size = 9.9,
391 https://www.jax.org/strain/009376). The J:ARC population is derived from an outbred stock
392 that has been subject to nearly a century of artificial selection for high fecundity, docility, and
393 other desirable laboratory traits including consistent body weight which is preferred for dosing
394 studies. In contrast, the J:DO are derived from phenotypically diverse recombinant inbred lines
395 (Collaborative Cross) and artificial selection for any phenotype is actively avoided through
396 randomized breeding that optimized for genetic diversity. These differences in selective
397 pressure may also explain the apparent loss of deleterious alleles through drift in the J:ARC.
398 That deleterious alleles have not been subject to purifying selection in the J:DO, also highlights
399 the potential buffering effects of high allelic diversity and heterozygosity.

400 Our data, together with previously published studies, reinforce the fact that not all
401 outbred stocks have comparable genetic and phenotypic diversity. Direct sequencing of
402 outbred stocks provides molecular basis for genetic quality control, breeding paradigms, and
403 optimization of experimental design for effective deployment these strains in biomedical
404 research.

405 *Study limitations.* Our exome sequencing data provide just a subset of the overall genetic
406 variation segregating in these outbred populations, specifically SNPs and small insertions /
407 deletions in coding sequence. Structural variation (SV), which we have not profiled here, is also
408 a significant source of genetic variation. SV have the potential to impact multiple genes with
409 potentially large phenotypic effects. While certain types of structural variants (deletions) can be
410 robustly detected using short read sequencing data, long read sequencing and *de novo*
411 assembly are the gold standard for genome-wide detection of SV. Moreover, outside of splice
412 junction sites and UTRs, our data do not capture non-coding variation, yet non-coding
413 variation is also a recognized source of phenotypic variation, especially for complex traits.
414 Finally, while our sequencing depth per sample is high, our samples sizes are low, making our
415 study underpowered for detection of ultra-rare SNPs/INDELs.

416

417 Tables

418

Stock	Total Variants	Mouse Collaborative Cross	Diversity Outbred Mouse Genomes	SangerSNP_Indel_ May_2020	dbSNP	EVA	Novel	High Impact Novel Variants
J:DO	478011	468487	427929	452989	439682	439642	4308	797
J:ARC	135825	116849	109808	121395	115557	115561	5105	408

419
420
421 **Table 1.** Total variants (SNPs/INDELs) and total novel variants in J:DO and J:ARC stocks.
422 Comparison of the total variants to all variants in the major, publicly available mouse variant
423 resource revealed that 0.9% and 3.75% of the total variants found in J:DO and J:ARC
424 respectively are novel.
425

426 Figure Legends

427 **Figure 1.** The Poiley method used for the importation and breeding of the JAX Swiss Outbred
428 (J:ARC) population. G0 were live animals from The Animal Resources Centre (ARC) in Canning
429 Vale that were subsequently bred and rederived through IVF to create 32 distinct breeding lines
430 at The Jackson Laboratory which continue to be maintained according to the Poiley method to
431 minimize inbreeding.
432
433 **Figure 2.** Heterozygosity in J:ARC (green) and J:DO (blue) datasets**.** The proportion (**A**) of
434 heterozygous calls per samples and distribution of heterozygous calls (**B**) for J:ARC (green) and

435 J:DO (blue) datasets. The mean heterozygosity for J:ARC samples was 0.25 while in J:DO it
436 was 0.39. Samples from both datasets were joint called and this analysis included only those
437 variants that are were called in both datasets (120,864)
438
439 **Figure 3.** Allele Frequency in J:ARC (green) and J:DO (blue) datasets. Samples from both
440 datasets were joint called and this analysis was restricted to those variants found in keeping
441 both datasets (120,864). Allele frequency was calculated using Plink. (**A)** MAF distribution for
442 120,864 SNPs for J:ARC (green) and J:DO (blue). Monomorphic variants are not removed. (**B**)
443 Allele frequency at the variant level for J:ARC samples (mean 0.18, median 0.13). (**C)** Allele
444 Frequency for J:DO samples at the variant level (mean 0.28, median 0.28).
445
446 **Figure 4**. Consensus segments of homozygosity for J:ARC (green) and J:DO (blue). The
447 number of consensus segments of homozygosity in J:ARC samples was higher **(A)** than in
448 J:DO samples **(B)**. Plink's default parameters were used. Only runs of homozygosity containing
449 at least 100 SNPs, and of total length ≥ 1000 kilobases, are noted.
450
451 **Figure 5.** Interindividual variability, relatedness, and LD decay. **(A)** PCA plot of J:DO (blue) and
452 J:ARC (green) joint variant calls. **(B)** A dendrogram of the J:ARC samples (using all called
453 variants in the dataset) showing the samples clustering in known sibling relationships. **(C)**
454 Comparison of LD decay J:ARC samples (green) and J:DO samples (blue). Each curve was
455 plotted using the 95th percentile of r2 values for SNPs spaced up to 1Mb apart. Plink was used
456 to calculate r^2 for all pairs of autosomal SNPs called from joint genotyping in the J:ARC
457 samples (425,409) and the J:DO samples (117,429). SNPs that were missing in more than 5%
458 of the samples and that were monomorphic were removed.
459
460 **Figure 6.** Novel variant consequences. **(A)** Showing the SnpEff effect categories for novel
461 variants for J:DO (on the left) and J:ARC (on the right) **(B)** High-impact (SnpEff effect
462 categories) novel variants are enriched in J:DO (on the left) in comparison to J:ARC (on the
463 right)
464

465 ## Supplementary Material

466 **Supplementary Data 1.** J:DO and J:ARC samples used for the study.

467

468 **Supplementary Data 2.** All novel, protein coding variants in J:ARC and J:DO with functional
469 annotation and allele frequency. Functional annotation (DAVID Bioinformatics Resources)
470 clustering of genes harboring novel variants in J:ARC and J:DO.

471

472 **Figure S1.** The Jackson Laboratory (JAX) Nexflow pipeline for laboratory mouse whole exome
473 sequencing.

474

475 **Figure S2.** Exome sequencing coverage statistics. **(A)** The mean target coverage versus high
476 quality filtered reads showing mean target coverage was 162X (91% target covered at 30X) for
477 J:ARC samples (green) and 155X (90% target covered at 30X) for J:DO (blue) samples. **(B)**
478 Percent target bases at 30X versus high quality filtered reads showing 90% of the target
479 exome regions were covered by 30 or more reads in both sets of samples.

480

481 **Figure S3.** The proportion of heterozygous variants based on the subset of variant calls for
482 known SNPs within the Mouse Universal Genotyping Array probe set. There were 3007 exome
483 variant calls at positions that are included in the GigaMUGA probe set. These were used to
484 estimate the proportion of heterozygous calls per J:DO using the R package, R/qtl2.
485 Estimates are based on autosomes at the haplotype level, using genotype probabilities of the 8
486 J:DO parental inbred strain genotypes.

487

488 **Figure S4.** Distribution of body weight in J:ARC and J:DO mice. Jackson Laboratory data
489 (unpublished, Schile 34608) and publicly available datasets from the Mouse Phenome
490 Database ((Shorter8, Chesler4 (LOGAN *et al.* 2013; RECLA *et al.* 2014),
491 https://phenome.jax.org/)and Wright et al. (WRIGHT *et al.* 2022) were used to assess the
492 distribution of body weight in J:DO and J:ARC mice). Data include both males and females and
493 sub-sets of 60 were randomly sampled from larger datasets for comparison. Overall body
494 weight in the J:ARC is higher than J:DO (ANOVA, adjusted P value <0.0001), while the
495 distribution of body weight (range) is greater in J:DO.

496

Acknowledgements

This work was partially supported by The Jackson Laboratory and the Mouse Mutant Resource and Research Center at The Jackson Laboratory (NIH U42 OD010921, LGR). We gratefully acknowledge the contribution of the Genome Technologies Service at The Jackson Laboratory for expert assistance with exome library preparation and sequencing.

Author Contributions

Conceptualization	K.B, L.R., A.Sr.
Methodology	A.S., B.C.
Formal analysis	A.S., B.C.
Investigation	A.S., B.C.
Resources	C.P., E.S., A. Sh.
Data Curation	A.S., B.C.
Writing, Review & Editing	L.R., A.Sr., B.C., K.B.
Visualization	B.C., A.Sr., L.R.
Supervision	A.Sr., L.R.
Funding acquisition	L.R., K.B., A.Sr.

Data availability

All the sequence data used in this study has been submitted to NCBI Sequence Read Archive under the bioproject PRJNA835415 and study SUB11284953.

LITERATURE CITED

Aldinger, K. A., G. Sokoloff, D. M. Rosenberg, A. A. Palmer and K. J. Millen, 2009 Genetic variation and population substructure in outbred CD-1 mice: implications for genome-wide association studies. PLoS One 4**:** e4729.

Bult, C. J., J. A. Blake, C. L. Smith, J. A. Kadin, J. E. Richardson *et al.*, 2019 Mouse Genome Database (MGD) 2019. Nucleic Acids Res 47**:** D801-D806.

525 Cezard, T., F. Cunningham, S. E. Hunt, B. Koylass, N. Kumar *et al.*, 2022 The European
526 Variation Archive: a FAIR resource of genomic variation for all species. Nucleic Acids
527 Res 50**:** D1216-D1220.
528 Chang, P. L., E. Kopania, S. Keeble, B. A. J. Sarver, E. Larson *et al.*, 2017 Whole exome
529 sequencing of wild-derived inbred strains of mice improves power to link phenotype and
530 genotype. Mamm Genome 28**:** 416-425.
531 Chia, R., F. Achilli, M. F. Festing and E. M. Fisher, 2005 The origins and uses of mouse outbred
532 stocks. Nat Genet 37**:** 1181-1186.
533 Cingolani, P., A. Platts, L. Wang le, M. Coon, T. Nguyen *et al.*, 2012 A program for annotating
534 and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the
535 genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6**:** 80-92.
536 Cui, S., C. Chesson and R. Hope, 1993 Genetic variation within and between strains of outbred
537 Swiss mice. Lab Anim 27**:** 116-123.
538 DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework
539 for variation discovery and genotyping using next-generation DNA sequencing data. Nat
540 Genet 43**:** 491-498.
541 Festing, M., K. Kondo, R. Loosli, S. M. Poiley and A. Spiegel, 1972 International standardized
542 nomenclature for outbred stocks of laboratory animals. Z Versuchstierkd 14**:** 215-224.
543 Foster, Y., L. Dutoit, S. Grosser, N. Dussex, B. J. Foster *et al.*, 2021 Genomic signatures of
544 inbreeding in a critically endangered parrot, the kakapo. G3 (Bethesda) 11.
545 Gatti, D. M., K. L. Svenson, A. Shabalin, L. Y. Wu, W. Valdar *et al.*, 2014 Quantitative trait
546 locus mapping methods for diversity outbred mice. G3 (Bethesda) 4**:** 1623-1633.
547 Groen, A., and A. J. Lagerwerf, 1979 Genic heterogeneity and genetic monitoring of mouse
548 outbred stocks. Lab Anim 13**:** 81-85.
549 Keane, T. M., L. Goodstadt, P. Danecek, M. A. White, K. Wong *et al.*, 2011 Mouse genomic
550 variation and its effect on phenotypes and gene regulation. Nature 477**:** 289-294.
551 Kinghorn, B., and A. Kinghorn, 2021 Management of inbreeding and coancestry to target short-
552 term and long-term genetic gains. Proc. Assoc. Advmt. Anim. Breed. Genet. 24**:** 191-194.
553 Lilue, J., A. G. Doran, I. T. Fiddes, M. Abrudan, J. Armstrong *et al.*, 2018 Sixteen diverse
554 laboratory mouse reference genomes define strain-specific haplotypes and novel
555 functional loci. Nat Genet 50**:** 1574-1583.

556 Logan, R. W., R. F. Robledo, J. M. Recla, V. M. Philip, J. A. Bubier *et al.*, 2013 High-precision
557 genetic mapping of behavioral traits in the diversity outbred mouse population. Genes
558 Brain Behav 12**:** 424-437.
559 McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome
560 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA
561 sequencing data. Genome Res 20**:** 1297-1303.
562 Morgan, A. P., C. P. Fu, C. Y. Kao, C. E. Welsh, J. P. Didion *et al.*, 2015 The Mouse Universal
563 Genotyping Array: From Substrains to Subspecies. G3 (Bethesda) 6**:** 263-279.
564 Mouse Genome Sequencing, C., R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers *et al.*,
565 2002 Initial sequencing and comparative analysis of the mouse genome. Nature 420**:** 520-
566 562.
567 Nicod, J., R. W. Davies, N. Cai, C. Hassett, L. Goodstadt *et al.*, 2016 Genome-wide association
568 of multiple complex traits in outbred mice by ultra-low-coverage sequencing. Nat Genet
569 48**:** 912-918.
570 Poiley, S. M., 1960 A systematic method of breeder rotation for non-inbred laboratory animal
571 colonies. Proc. Anim. Care Panel 10**:** 159-166.
572 Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira *et al.*, 2007 PLINK: a tool set
573 for whole-genome association and population-based linkage analyses. Am J Hum Genet
574 81**:** 559-575.
575 Recla, J. M., R. F. Robledo, D. M. Gatti, C. J. Bult, G. A. Churchill *et al.*, 2014 Precise genetic
576 mapping and integrative bioinformatics in Diversity Outbred mice reveals Hydin as a
577 novel pain gene. Mamm Genome 25**:** 211-222.
578 Rice, M. C., and S. J. O'Brien, 1980 Genetic variance of laboratory outbred Swiss mice. Nature
579 283**:** 157-161.
580 Saul, M. C., V. M. Philip, L. G. Reinholdt, A. Center for Systems Neurogenetics of and E. J.
581 Chesler, 2019 High-Diversity Mouse Populations for Complex Traits. Trends Genet 35**:**
582 501-514.
583 Sherman, B. T., M. Hao, J. Qiu, X. Jiao, M. W. Baseler *et al.*, 2022 DAVID: a web server for
584 functional enrichment analysis and functional annotation of gene lists (2021 update).
585 Nucleic Acids Res.

586 Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan *et al.*, 2001 dbSNP: the NCBI
587 database of genetic variation. Nucleic Acids Res 29**:** 308-311.
588 Sigmon, J. S., M. W. Blanchard, R. S. Baric, T. A. Bell, J. Brennan *et al.*, 2020 Content and
589 Performance of the MiniMUGA Genotyping Array: A New Tool To Improve Rigor and
590 Reproducibility in Mouse Research. Genetics 216**:** 905-930.
591 Srivastava, A., A. P. Morgan, M. L. Najarian, V. K. Sarsani, J. S. Sigmon *et al.*, 2017 Genomes
592 of the Mouse Collaborative Cross. Genetics 206**:** 537-556.
593 Stoffel, M., M. Esser, M. Kardo, E. Humble, H. Nichols *et al.*, 2016 inbreedR: an R package for
594 the analysis of inbreeding based on genetic markers. Methods in Ecology and Evolution
595 7**:** 1331-1339.
596 Strobel, M. C., L. G. Reinholdt, R. D. Malcolm and K. Pritchett-Corning, 2015 Genetic
597 Monitoring of Laboratory Rats and Mice, pp. 1403-1416 in *Laboratory Animal Medicine,*
598 *3rd Edition*, edited by J. G. Fox, L. C. Anderson, G. Otto, K. R. Pritchett-Corning and M.
599 T. Whary. Elsevier, Boston.
600 Weichman, K., and J. R. Chaillet, 1997 Phenotypic variation in a genetically identical population
601 of mice. Mol Cell Biol 17**:** 5269-5274.
602 Wright, K. M., A. G. Deighan, A. Di Francesco, A. Freund, V. Jojic *et al.*, 2022 Age and diet
603 shape the genetic architecture of body weight in diversity outbred mice. Elife 11.
604 Yalcin, B., J. Nicod, A. Bhomra, S. Davidson, J. Cleak *et al.*, 2010 Commercially available
605 outbred mice for genome-wide association studies. PLoS Genet 6**:** e1001085.
606 Yalcin, B., K. Wong, A. Agam, M. Goodson, T. M. Keane *et al.*, 2011 Sequence-based
607 characterization of structural variation in the mouse genome. Nature 477**:** 326-329.
608

FIGURE 1

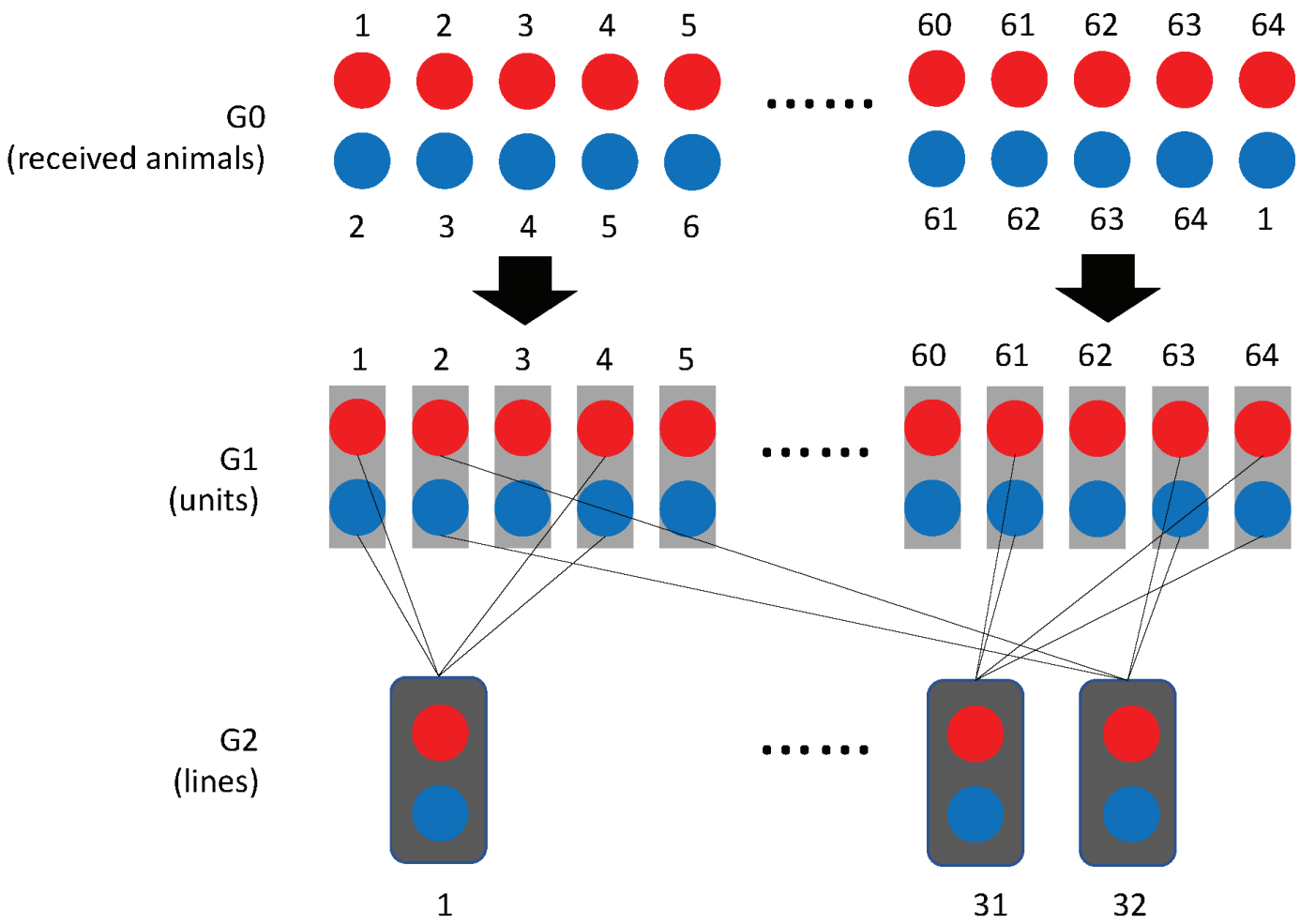

G0
(received animals)
1 2 3 4 5 60 61 62 63 64
2 3 4 5 6 61 62 63 64 1
G1
(units)
1 2 3 4 5 60 61 62 63 64
G2
(lines)
1 31 32

FIGURE 2

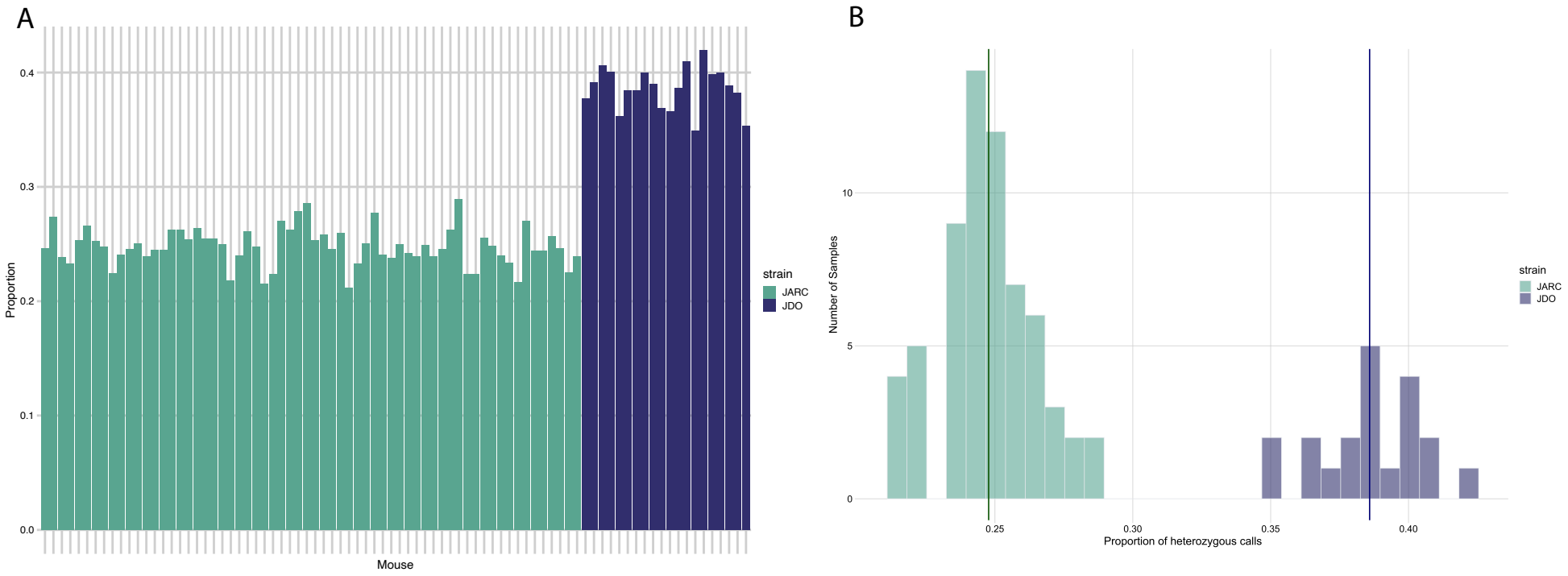

A
Proportion
0.0
0.1
0.2
0.3
0.4
Mouse
strain
JARC
JDO
B
Number of Samples
0
5
10
0.25
0.30
0.35
0.40
Proportion of heterozygous calls
strain
JARC
JDO

FIGURE 3

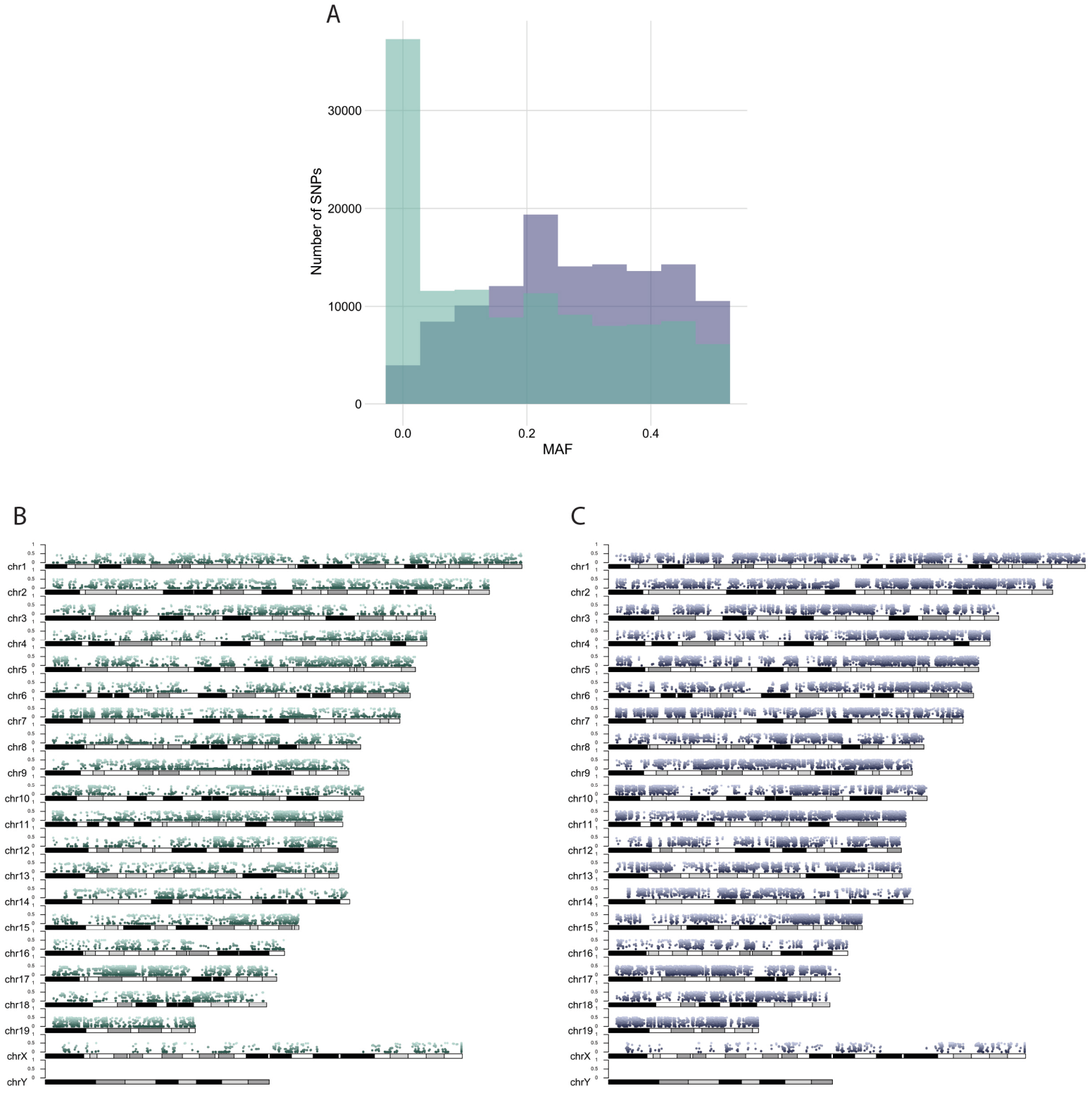

A
Number of SNPs
30000
20000
10000
0
0.0
0.2
0.4
MAF
B
C
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chrX
chrY

FIGURE 4

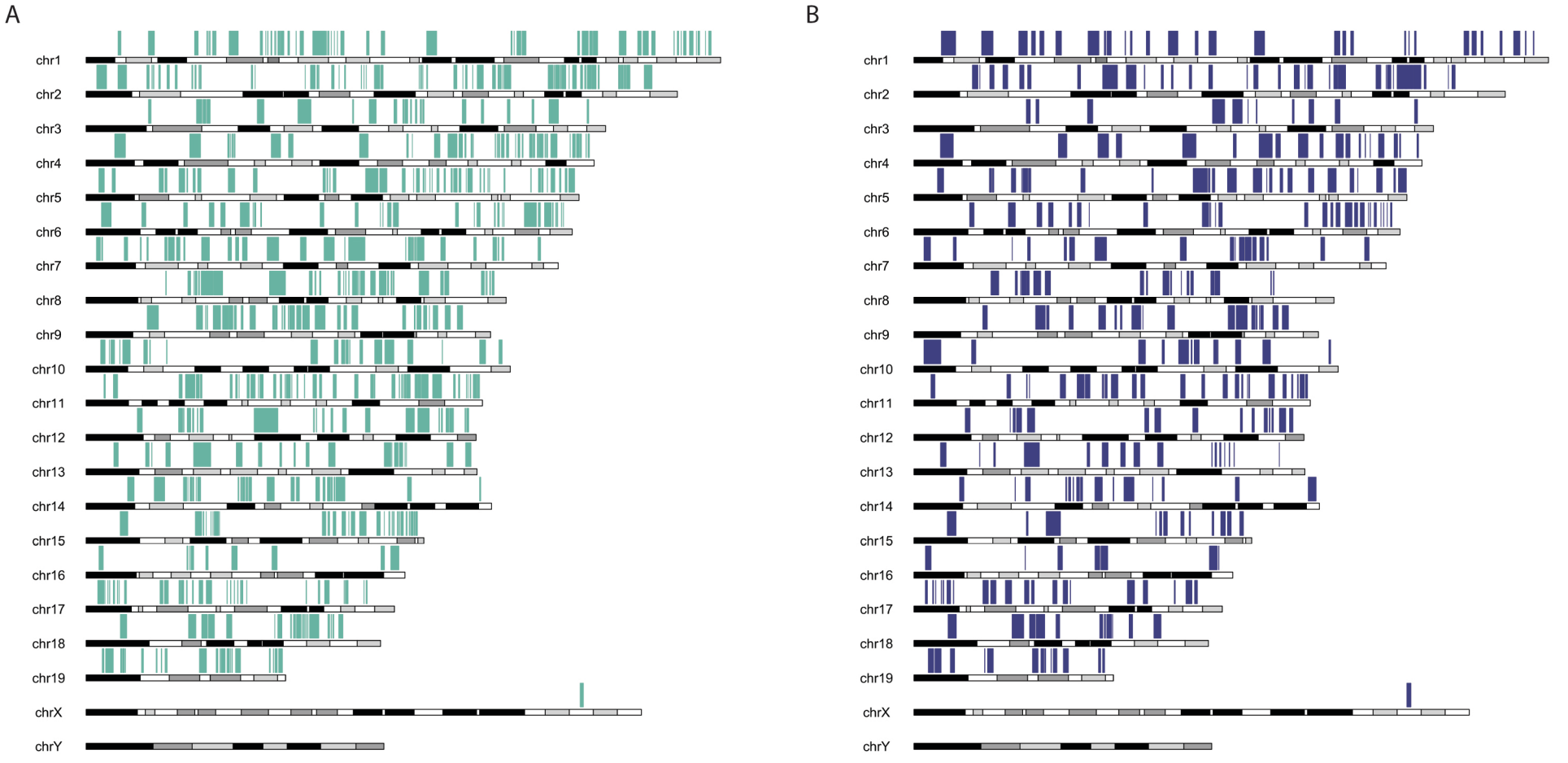
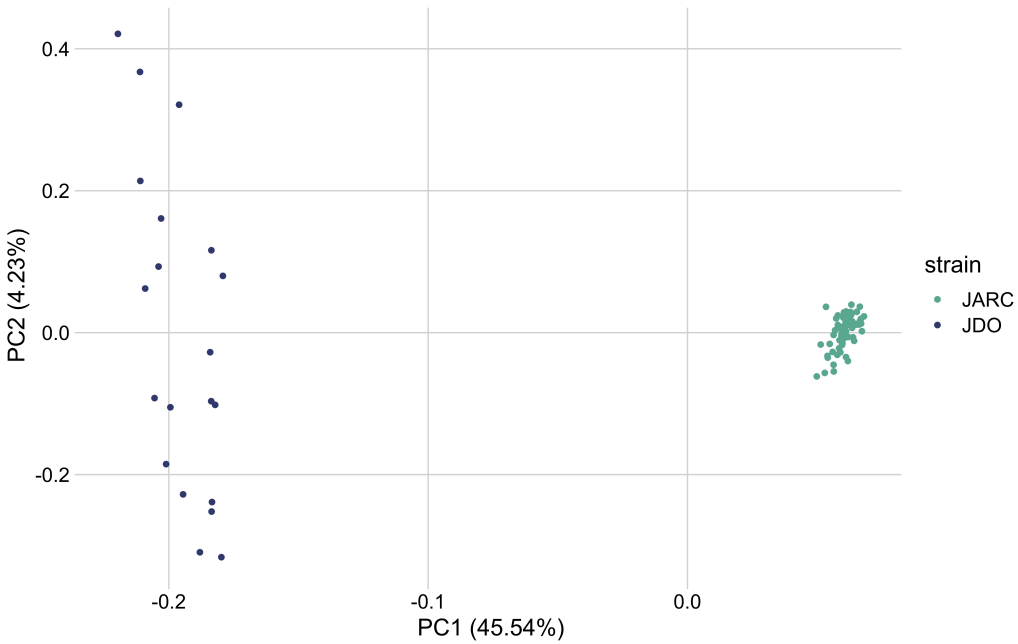
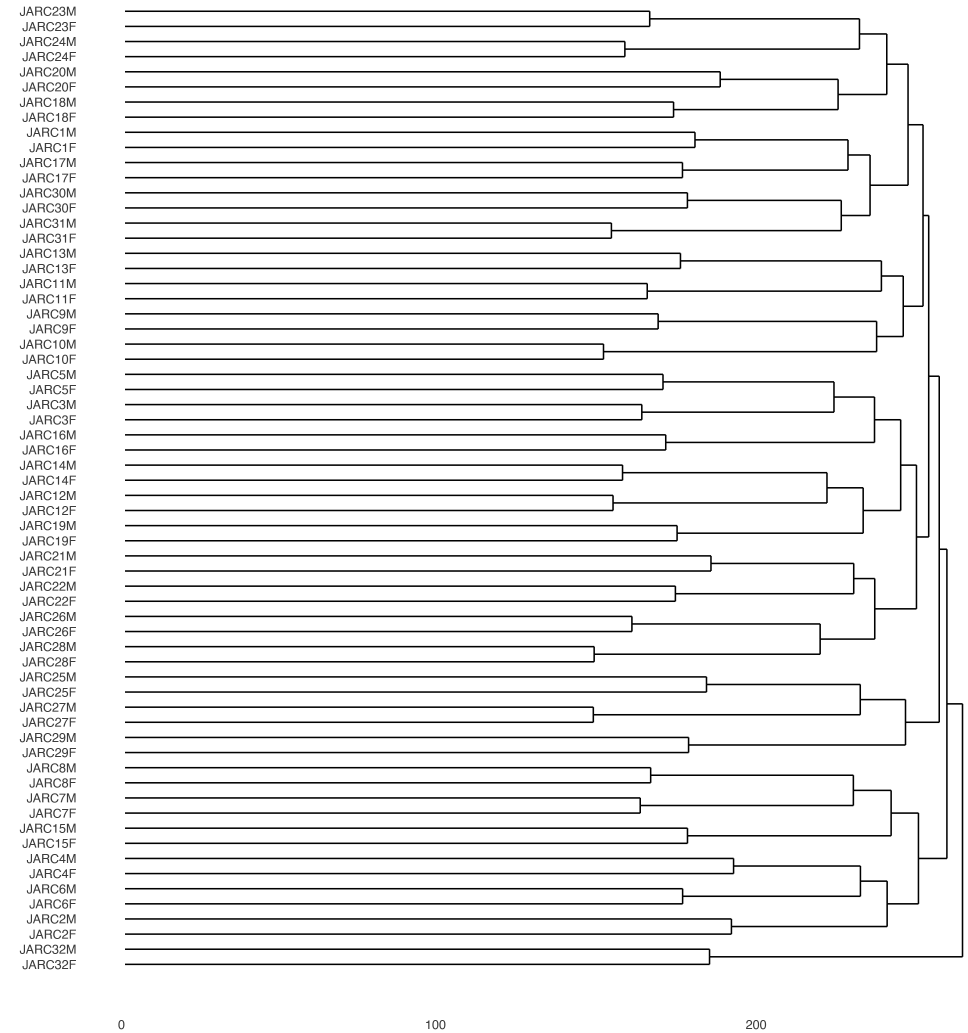

A
B
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chrX
chrY

FIGURE 5

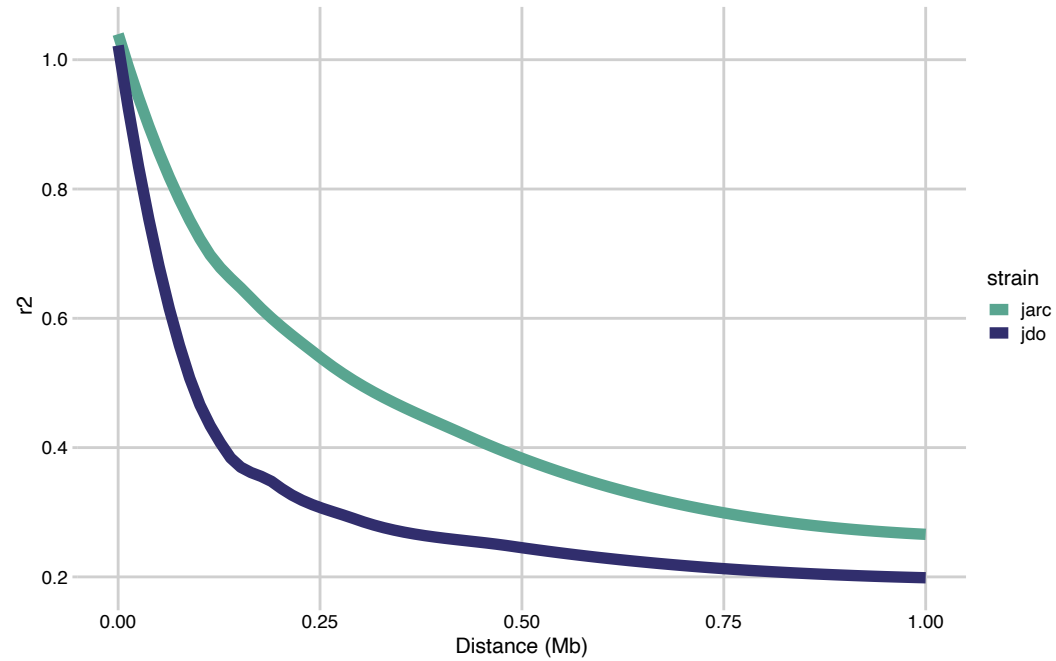
A

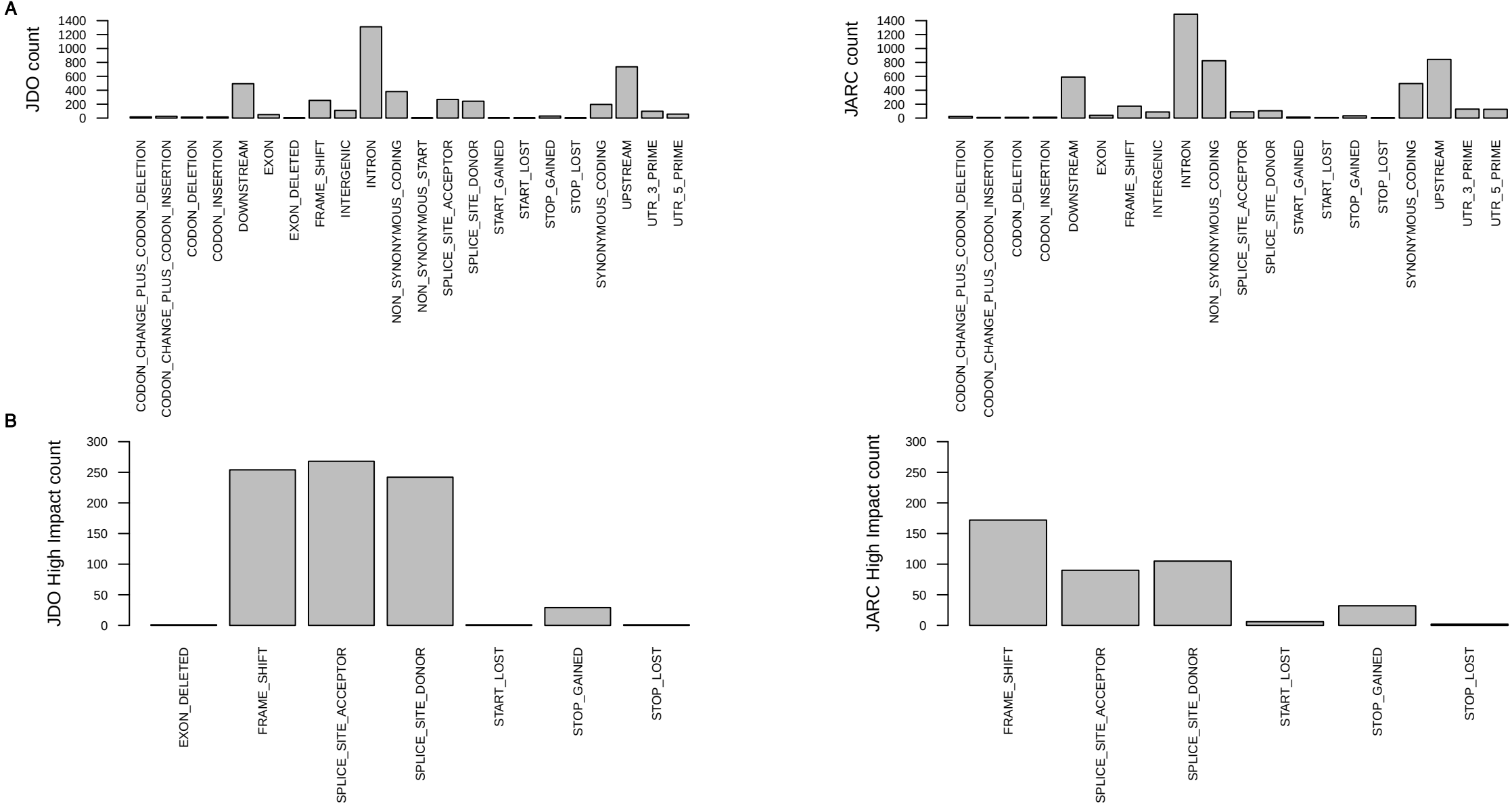

0.4
0.2
0.0
-0.2
PC2 (4.23%)
-0.2
-0.1
0.0
PC1 (45.54%)
strain
JARC
JDO

B

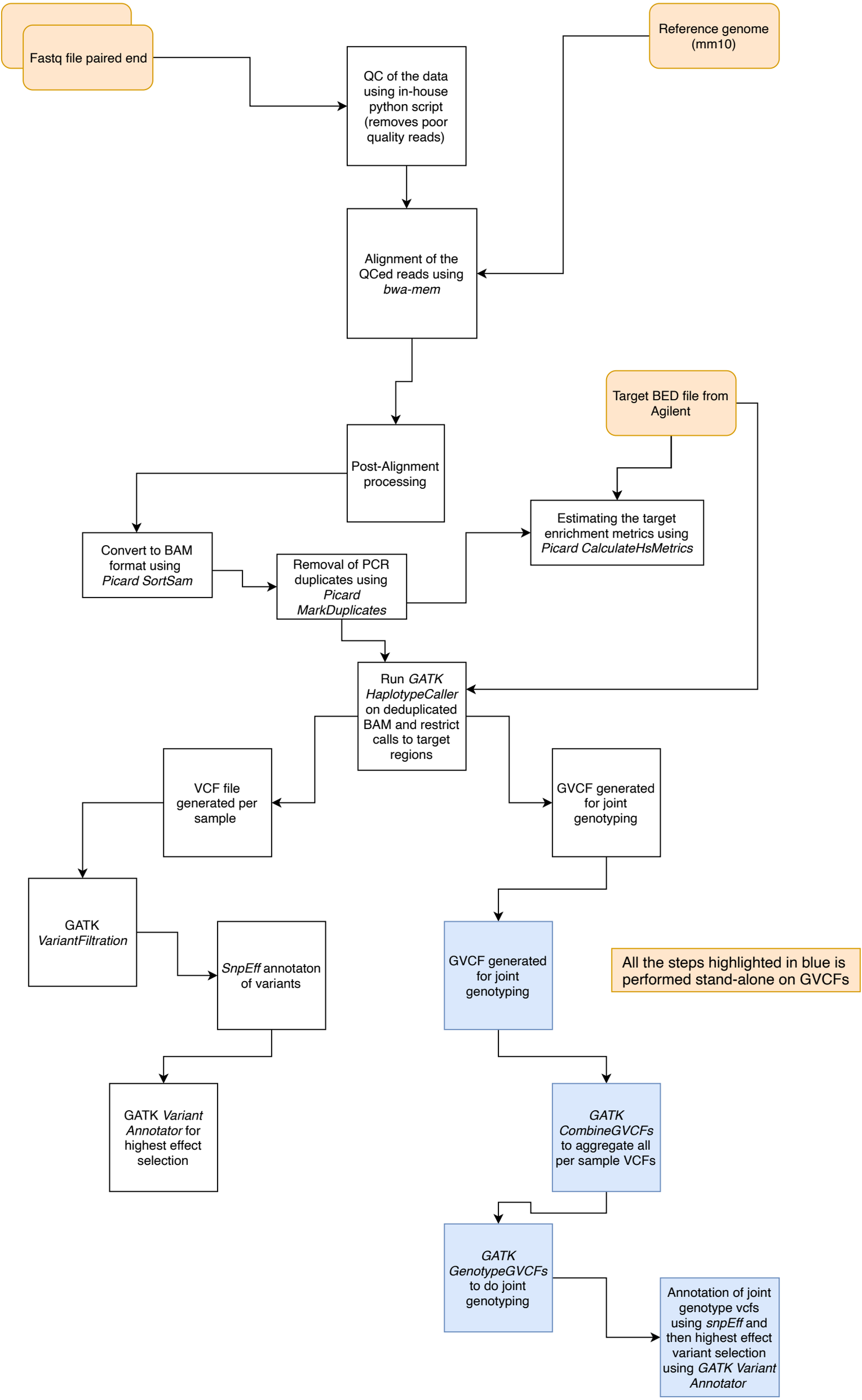

JARC23M
JARC23F
JARC24M
JARC24F
JARC20M
JARC20F
JARC18M
JARC18F
JARC1M
JARC1F
JARC17M
JARC17F
JARC30M
JARC30F
JARC31M
JARC31F
JARC13M
JARC13F
JARC11M
JARC11F
JARC9M
JARC9F
JARC10M
JARC10F
JARC5M
JARC5F
JARC3M
JARC3F
JARC16M
JARC16F
JARC14M
JARC14F
JARC12M
JARC12F
JARC19M
JARC19F
JARC21M
JARC21F
JARC22M
JARC22F
JARC26M
JARC26F
JARC28M
JARC28F
JARC25M
JARC25F
JARC27M
JARC27F
JARC29M
JARC29F
JARC8M
JARC8F
JARC7M
JARC7F
JARC15M
JARC15F
JARC4M
JARC4F
JARC6M
JARC6F
JARC2M
JARC2F
JARC32M
JARC32F
0
100
200

C


1.0
0.8
0.6
0.4
0.2
r2
0.00
0.25
0.50
0.75
1.00
Distance (Mb)
strain
jarc
jdo

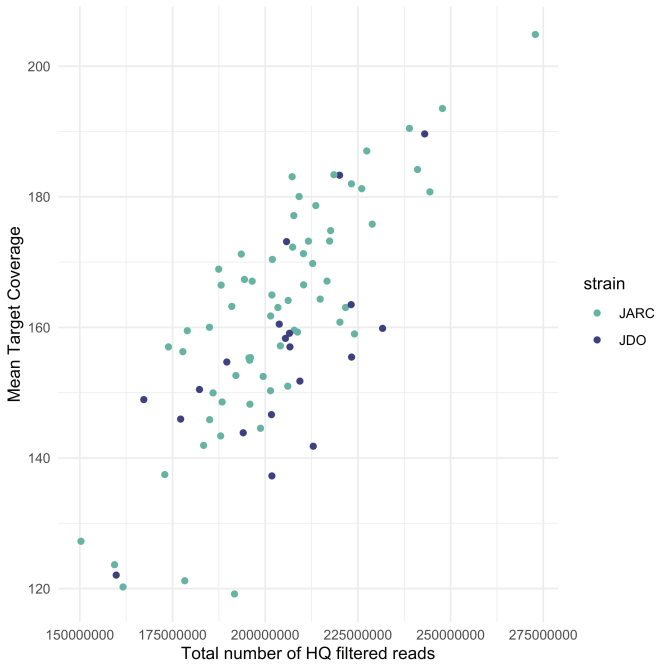

A
JDO count
1400
1200
1000
800
600
400
200
0
CODON_CHANGE_PLUS_CODON_DELETION
CODON_CHANGE_PLUS_CODON_INSERTION
CODON_DELETION
CODON_INSERTION
DOWNSTREAM
EXON
EXON_DELETED
FRAME_SHIFT
INTERGENIC
INTRON
NON_SYNONYMOUS_CODING
NON_SYNONYMOUS_START
SPLICE_SITE_ACCEPTOR
SPLICE_SITE_DONOR
START_GAINED
START_LOST
STOP_GAINED
STOP_LOST
SYNONYMOUS_CODING
UPSTREAM
UTR_3_PRIME
UTR_5_PRIME
JARC count
1400
1200
1000
800
600
400
200
0
CODON_CHANGE_PLUS_CODON_DELETION
CODON_CHANGE_PLUS_CODON_INSERTION
CODON_DELETION
CODON_INSERTION
DOWNSTREAM
EXON
FRAME_SHIFT
INTERGENIC
INTRON
NON_SYNONYMOUS_CODING
SPLICE_SITE_ACCEPTOR
SPLICE_SITE_DONOR
START_GAINED
START_LOST
STOP_GAINED
STOP_LOST
SYNONYMOUS_CODING
UPSTREAM
UTR_3_PRIME
UTR_5_PRIME
B
JDO High Impact count
300
250
200
150
100
50
0
EXON_DELETED
FRAME_SHIFT
SPLICE_SITE_ACCEPTOR
SPLICE_SITE_DONOR
START_LOST
STOP_GAINED
STOP_LOST
JARC High Impact count
300
250
200
150
100
50
0
FRAME_SHIFT
SPLICE_SITE_ACCEPTOR
SPLICE_SITE_DONOR
START_LOST
STOP_GAINED
STOP_LOST

SUPPLEMENTARY FIGURE S1

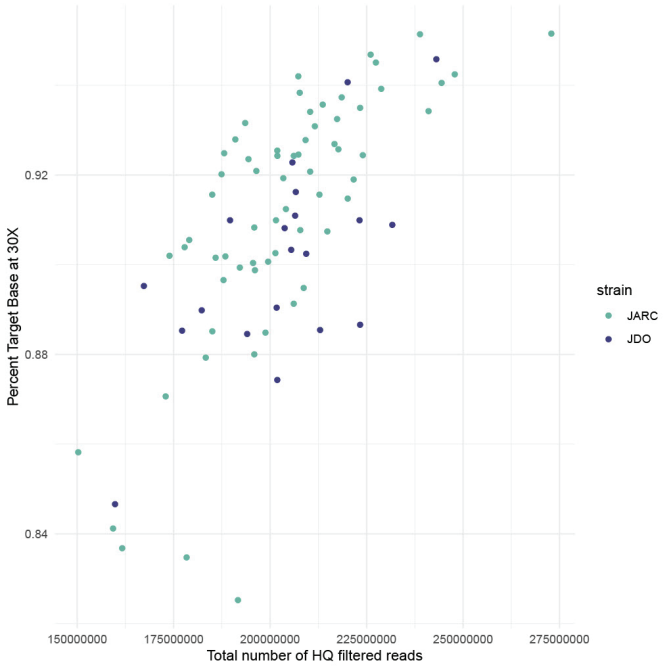


SUPPLEMENTARY FIGURE S2

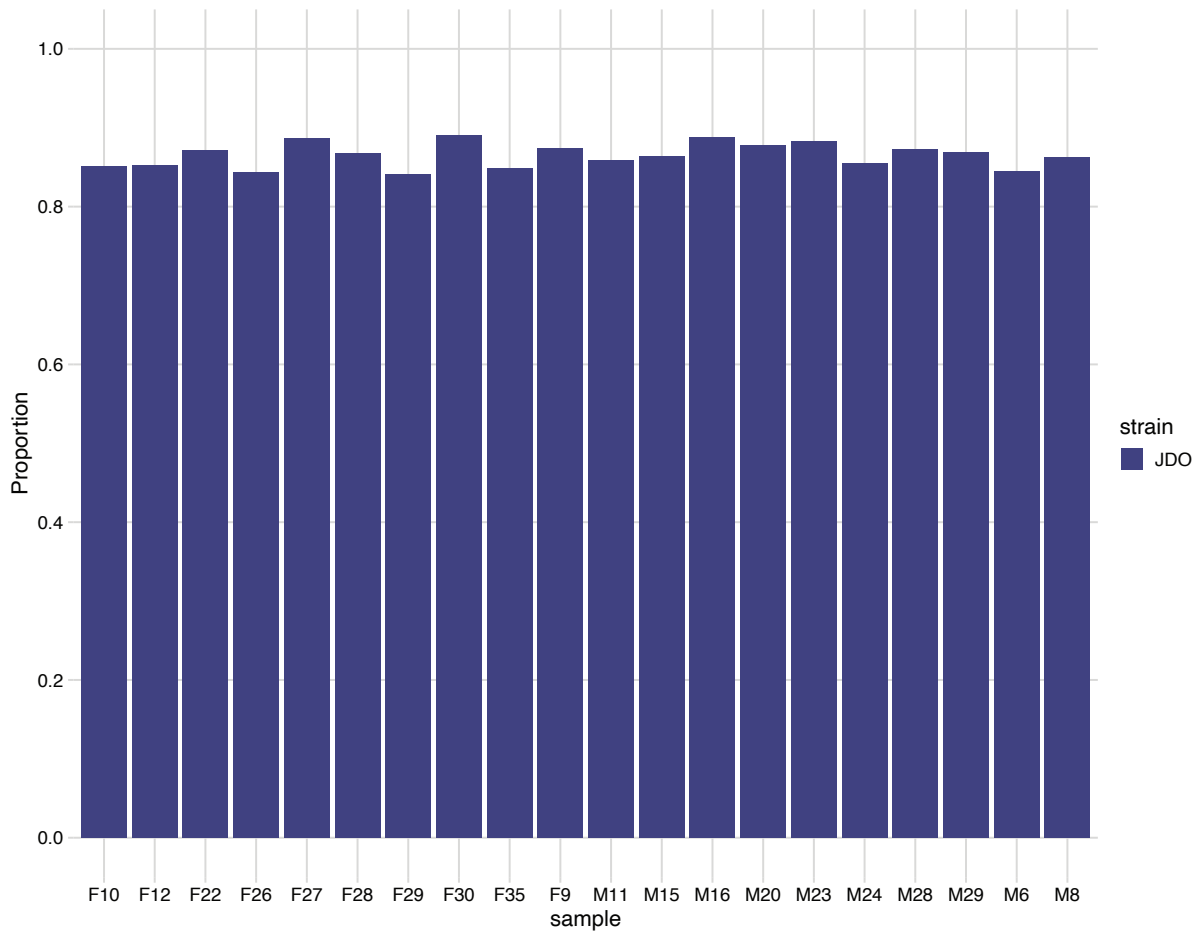
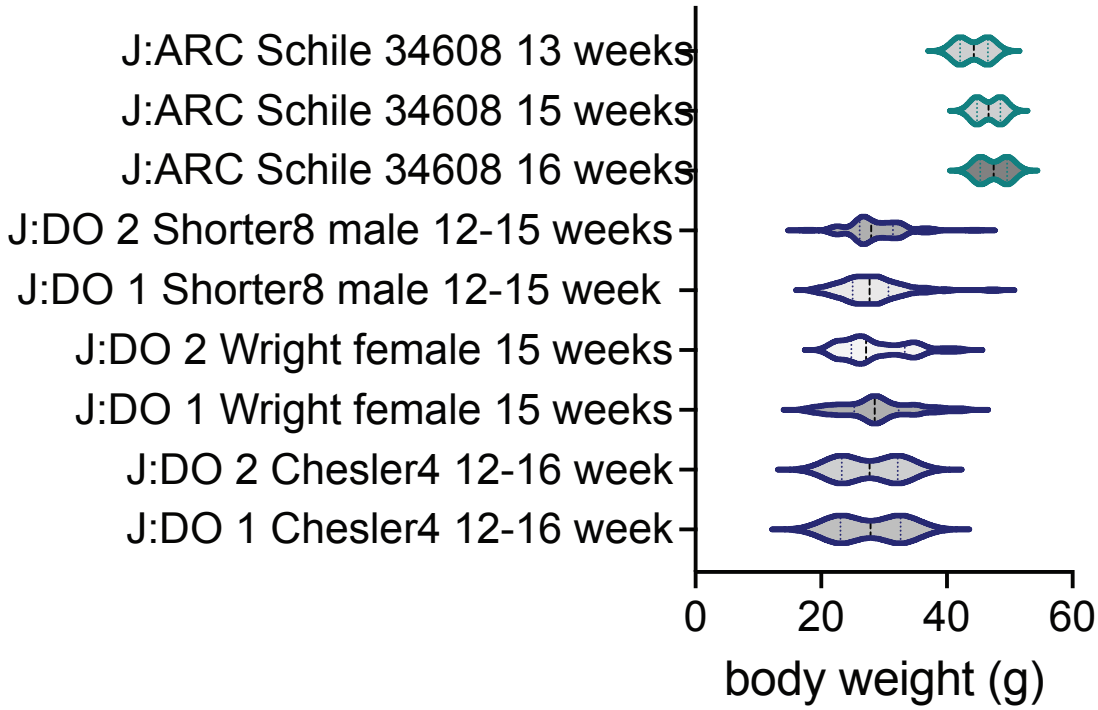
A



B



SUPPLEMENTARY FIGURE S3



SUPPLEMENTARY FIGURE S4