Abundant genetic variation is retained in many laboratory schistosome populations

Authors: Kathrin S. JUTZELER[1,2*], Roy N. PLATT[3], Robbie DIAZ[4], Madison MORALES[4], Winka LE CLEC'H[1], Frédéric D. CHEVALIER[1], Timothy J.C. ANDERSON[3*]

Affiliation:

[1] Host parasite Interaction Program, Texas Biomedical Research Institute, P.O. Box 760549, 78245 San Antonio, Texas, USA.

[2] UT Health, Microbiology, Immunology & Molecular Genetics, San Antonio, TX 78229

[3] Disease Intervention and Prevention program, Texas Biomedical Research Institute, P.O. Box 760549, 78245 San Antonio, Texas, USA.

[4] Texas Biomedical Research Institute, P.O. Box 760549, 78245 San Antonio, Texas, USA.

Email addresses:

Kathrin S. JUTZELER: kjutzeler@txbiomed.org – ORCID: 0000-0002-3687-4020

Roy N. PLATT: rplatt@txbiomed.org – ORCID: 0000-0002-9754-5765

Winka LE CLEC'H: winkal@txbiomed.org – ORCID: 0000-0002-1111-2492

Frédéric D. CHEVALIER: fcheval@txbiomed.org – ORCID : 0000-0003-2611-8106

Timothy J.C. ANDERSON: tanderso@txbiomed.org – ORCID: 0000-0002-0191-0204

*Corresponding authors

ABSTRACT:

Schistosomes are obligately sexual blood flukes that can be maintained in the laboratory using freshwater snails as intermediate and rodents as definitive hosts. The genetic composition of laboratory schistosome populations is poorly understood: whether genetic variation has been purged due to serial inbreeding or retained is unclear. We sequenced 19 – 24 parasites from each of five laboratory *Schistosoma mansoni* populations and compared their genomes with published exome data from four *S. mansoni* field populations. We found abundant genomic variation (0.897 – 1.22 million variants) within laboratory populations: these retained on average 49% (π = 3.27e-04 – 8.94e-04) of the nucleotide diversity observed in the four field parasite populations (π = 1.08e-03 – 2.2e-03). However, the pattern of variation was very different in laboratory and field populations. Tajima's D was positive in all laboratory populations except SmBRE, indicative of recent population bottlenecks, but negative in all field populations. Current effective population size estimates of laboratory populations were lower (2 – 258) compared to field populations (3,174 – infinity). The distance between markers at which linkage disequilibrium (LD) decayed to 0.5 was longer in laboratory populations (59 bp – 180 kb) compared to field populations (9 bp – 9.5 kb). SmBRE was the least variable; this parasite also shows low fitness across the lifecycle, consistent with inbreeding depression. The abundant genetic variation present in most laboratory schistosome populations has several important implications: (i) measurement of parasite phenotypes, such as drug resistance, using laboratory parasite populations will determine average values and underestimate trait variation; (ii) genome-wide association studies (GWAS) can be conducted in laboratory schistosome populations by measuring phenotypes and genotypes of individual worms; (iii) genetic drift may lead to divergence in schistosome populations maintained in different laboratories. We conclude that the abundant genetic variation retained within

41 many laboratory schistosome populations can provide valuable, untapped opportunities for

42 schistosome research.

43

44 **KEY WORDS:** *Schistosoma mansoni*, genomic variation, genetic diversity, effective population size,

45 linkage disequilibrium, genome-wide association studies (GWAS)

46

BACKGROUND

Many viral, bacterial and protozoan pathogens can be cloned and maintained as asexual lineages in the laboratory. This has many advantages for research because experimental infections can be established using genetically homogeneous pathogens, and differences in biomedically important pathogen traits can be directly attributed to genetic differences between pathogen clones. In contrast, the blood fluke *Schistosoma mansoni* has separate sexes (males are ZZ; females are ZW) and an obligately sexual reproductive system: these parasites are maintained as recombining populations in the laboratory. Successful cryopreservation has been reported for schistosomes but is inconsistent [1], and cannot be used reliably for maintaining schistosome populations. Schistosome populations are therefore typically maintained by continuous passage through the aquatic snail intermediate host, where clonal proliferation of larval stages occurs, and the rodent definitive host, where adult males and females pair and produce eggs.

Schistosome populations have been maintained in the laboratory for up to 80 years [2]. For example, the SmNMRI parasite population maintained by the Biomedical Research Institute (BRI) [3] was originally isolated in the 1940s [2]. Our laboratory maintains four different parasite populations: SmEG from Egypt, collected at an undetermined date (possibly in the 1980s) by US researchers and then established at the Theodor Bilharz Research Institute in Cairo in 1990 [4,5]. SmLE isolated in Brazil in 1965 [2], while SmBRE was acquired from Brazil in 1975 [6], and SmOR, a descendant from SmHR, which was isolated in Puerto Rico in 1971 [7]. Assuming five generations per year, these parasite populations have been maintained continuously for ~400 (SmNMRI), ~160 (SmEG), ~285 (SmLE), ~235 (SmBRE), and 270 (SmOR) generations.

The genomic consequences of long-term laboratory passage in schistosomes are not known, but several authors investigated this question in the pre-genomic era. Fletcher et al. [8] examined enzyme polymorphism at 18 loci in individual worms. They measured mean heterozygosity per locus and observed that genetic variation within laboratory populations maintained from 1-40 generations was approximately half that observed in fresh parasite isolates. Minchella et al. [9] quantified genetic variation in a maternally inherited DNA element (pSM750) using restriction fragment length polymorphism (RFLP) of individual parasites from 14 laboratory isolates. They noted that parasites from the same laboratory isolate generally showed low variability. However, SmNMRI parasites exhibited extensive variation. Pinto et al. [10] found no variation between worms from a laboratory isolate (SmLE), but extensive variation within parasites derived from different Brazilian patients using random amplified polymorphic DNA (RAPD) analysis from three different primer sets. Hence, these studies reached rather different conclusions.

Efforts to sequence the genome of *S. mansoni* provided further insights. The *S. mansoni* genome was initially sequenced from pools of parasites from the SmNMRI population [11]. The genetic variation present within these populations contributed to issues with genome assembly: the resultant assembly was fragmented in > 19,000 scaffolds [11,12]. As a consequence, subsequent work to improve the genome used DNA isolated from worms with a single genotype, that were a product of single miracidium larvae infections, to minimize this issue. This approach contributed to a much improved genome assembly, closing more than 40,000 gaps and assigning 81% of the data to chromosomes [13].

Phenotypic data provides further evidence that parasite populations may not be homogeneous. Davies et al. isolated parasites that shed low or high numbers of cercariae from the SmPR population [14]. Furthermore, they were able to select low and high shedding populations [15], indicating this

phenotypic variation has a genetic basis. Similarly Le Clec'h et al. [16] demonstrated that the SmLE-PZQ-R population, which was selected for resistance to praziquantel (PZQ) in the SmLE population from Brazil, comprises a mixture of praziquantel (PZQ) resistant and sensitive parasites, as well as abundant variation across the genome [17].

This study was designed to directly measure genomic variation within five laboratory schistosome populations. We speculated that either i) a low number of founders or inbreeding due to repeated laboratory passage could result in bottlenecks and therefore a loss of genetic variation or ii) sexual outbreeding could be sufficient to retain high levels of genetic variation (Figure 1). We generated 117 independent genome sequences from four schistosome populations maintained in our laboratory and from the widely used SmNMRI population maintained at the BRI. We compared variation in these laboratory populations with published exome sequence data from field collected *S. mansoni* parasites from Brazil, Niger, Senegal, and Tanzania [18]. We observed abundant genetic variation within laboratory populations, albeit reduced by 51% compared to field collected parasites. However, laboratory and field collected parasites showed dramatic differences in pattern of variation, including the allele frequency spectrum, linkage disequilibrium, and effective population size (N_e). We evaluate the implications of these results for schistosome research.

106 **RESULTS**

107 **Summary of sequence data**

108 We sequenced the genomes of 117 *S. mansoni* parasites from five populations (Fig. 2). We retained 108
109 of 117 generated genome sequences from laboratory samples after quality filtering: 19 SmNMRI, 21
110 SmOR, 20 SmBRE, and 24 each for SmEG and SmLE. The mean read depths for these samples was 32.8x
111 (range: 10.0 – 143.8x), and we discovered 0.897 – 1.22 million single nucleotide polymorphisms (SNPs)
112 in the laboratory populations. Detailed information about these variants is listed in Table 1. In addition,
113 we kept 124 previously generated exome sequences from Brazil (*n* = 46), Niger (*n* = 9), Senegal (*n* = 24),
114 and Tanzania (*n* = 45) [18]. To make field and laboratory samples directly comparable, we filtered
115 genotyped laboratory and field samples jointly, keeping only variants that fall in the coding sequence
116 (CDS) region. This resulted in 362,190 variants of which 281,680 were autosomal (Table 2). Coverage
117 statistics for each sample are listed in Table S1.

118

119 **Principal component analysis (PCA) and admixture**

120 We generated a PCA plot using 1.24 million MAF filtered, autosomal variants (MAF > 0.05) from our
121 laboratory genome sequences (Figure 3A). This analysis identified five distinct clusters. While SmOR,
122 SmEG, SmNMRI, and SmLE all clustered along the vertical axis, SmBRE formed a separate cluster along
123 the horizontal axis.

We used ADMIXTURE and plotted five populations, as k = 5 resulted in the smallest cross-validation score (Figure 3B). This analysis confirmed the presence of five schistosome populations with distinct allelic components.

Nucleotide diversity in *S. mansoni* laboratory and field populations

The distribution of SNP variation across the genome is shown in Figure 4A. We calculated nucleotide diversity (π) in 25 kb windows (Figure 4B). Statistical analysis using a Kolmogorov-Smirnov test showed a significant reduction in diversity (51%) in laboratory populations (D = 0.403, $p < 0.001$). As previously documented, samples from Tanzania had the highest nucleotide diversity of all populations [18]. This analysis also revealed minimal diversity in the SmBRE population. While SmBRE had 1.26E+05 segregating SNPs (MAF > 0.05), equivalent numbers for the other populations were 8.69E+05 (SmEG), 5.23E+05 (SmLE), 6.40E+05 (SmOR) and 7.23E+05 (SmNMRI) (Table 1).

Tajima's D and allele frequency distributions

Tajima's D revealed a stark contrast between laboratory-maintained and field populations: while four of five laboratory populations exhibited a positive Tajima's D, all field populations showed negative values (Figure 5A; Wilcoxon test; W = 17, $p = 0.111$). The exception to this was SmBRE, which also had a negative Tajima's D like the field populations. We inspected allele frequency spectra in each population. This revealed SNPs at intermediate frequencies were common in SmEG, SmLE, SmOR, and SmNMRI, whereas field populations (and SmBRE) had a high frequency of rare alleles (Figure S1). We plotted the empirical

144 cumulative distribution (ECDF) of allele frequencies for each population (Figure 5B), revealing highly
145 significant differences between allele frequency spectra for laboratory and field populations (two sample
146 Kolmogorov-Smirnov test: D = 0.486, $p < 2.2e-16$).

147

148 **Linkage disequilibrium in laboratory and field populations**

149 We calculated linkage disequilibrium (LD) for each *S. mansoni* population and estimated LD decay with
150 physical distance between markers from pairwise r^2 values. As we only retained 1,215 common (MAF >
151 0.05) exonic SNPs in the SmBRE population, we used all autosomal variants to calculate LD decay in the
152 laboratory populations. Figure 6A shows slower LD decay in four out of the five laboratory populations
153 compared to the field populations. To compare LD decay curves, we measured the distance at which LD
154 is reduced to $r^2 = 0.5$ ($LD_{0.5}$, Figure 6B). LD decayed extremely rapidly in the Tanzanian parasite population
155 ($LD_{0.5 =}$ 9 bp). LD decayed uniformly in the Nigerien, Senegalese, and Brazilian populations, with $LD_{0.5}$
156 ranging from 1,000 to 9,543 bp. LD decay was nearly significantly slower in the laboratory populations
157 (T-test, $t_{(6.23)} = 2.31$, p = 0.058), with $LD_{0.5}$ ranging from 72 kb to 180 kb in SmEG, SmLE, SmNMRI, and
158 SmOR. In stark contrast to other laboratory populations, SmBRE exhibited very rapid LD decay ($LD_{0.5}$ =
159 59 bp). We also calculated LD using exonic SNPs only to ensure that the differences observed did not
160 result from use of different marker sets in field and laboratory populations (Figure S3). This confirmed
161 slower LD decay in laboratory than field populations (T-test, $t_{(4.04)} = 3.23$, $p = 0.032$), with the exception
162 of SmBRE.

Population size

We used our sequencing data to predict the current effective population size (N_e) based on either linkage disequilibrium (NeEstimator) or sibship frequency (COLONY). NeEstimator computed effective population sizes ranging from 2 – 258 in the laboratory and 3,174 (Brazil) – infinity (Niger, Senegal, Tanzania) in the field populations (Figure 7A), while COLONY reported N_e values from 5 – 123 for laboratory populations and 3,612 – infinity for field populations (Figure 7B). Both NeEstimator and COLONY identified SmNMRI and SmLE as having the highest N_e estimates among the laboratory populations, while SmBRE had the lowest N_e. N_e estimates for laboratory populations using both approaches were correlated ($R^2 = 0.96$, $p = 0.020$). N_e estimates were at least 12-fold greater in field than in laboratory schistosome populations with NeEstimator and at least 29-fold greater with COLONY.

Using our life cycle maintenance records, we estimated the census size (N_c) of our four laboratory schistosome populations over time and calculated the harmonic mean of each population. This was done by estimating the number of parasite genotypes used to infect hamsters for each laboratory maintenance cycle over a seven-year period (Figure S2). We did not have census data for the SmNMRI population maintained at BRI. Census size remained relatively consistent in SmLE, SmOR, and SmEG. However, population size increased in SmBRE parasites starting in 2021 (Figure S2A). The reasons for this are explained elsewhere [19]. SmLE had the highest census with 157 genotypes, followed by SmOR (137) and SmEG (132), and SmBRE (93) (Figure S2). Population size data is summarized in Table 3.

Simulations of genomic diversity in populations of different size

We conducted simulations to examine how population size impacts retention of diversity in schistosomes. When N = 5, simulations show that autosomal diversity is reduced by >99% in 55 generations when the life cycle is maintained with overlapping generations and 91 generations without overlap. When N = 100, we observe an 82.7% reduction in diversity with overlap and a 63.4% reduction without overlap relative to the progenitor population after 400 generations. When N = 200, the reduction is 59.9% and 39.9% respectively. The average effective population size (N_e) in our laboratory populations is 124, as calculated by NeEstimator, and 68, as calculated using Colony.

DISCUSSION

High levels of genetic diversity in most laboratory schistosome populations

We sequenced parasites from five different laboratory-maintained *S. mansoni* populations and compared them to four field populations from Africa and South America. Our genomic data revealed 0.897 – 1.22 million variants segregating within the five laboratory populations. This is equivalent to one variant every 321- 436 bp. Furthermore, our study revealed 51% lower nucleotide diversity (π) in exome data from laboratory-maintained schistosome populations than from field populations. Despite repeated passage over 30 – 80 years (~150-400 generations, assuming five generations per year), only half of genetic diversity is lost in laboratory schistosome populations.

Other studies have compared the genetic composition of different wild and domesticated/farmed species. Domestication of animals or plants often involves a low number of founders and the selection for specific traits [20–24]. Nucleotide diversity (π) is reduced in domesticated species populations by between 33 and 98% relative to wild populations (Table 4) [25,26]. The notable exception is for pigs, in which some populations show greater diversity than wild boar [27,28]. Laboratory *S. mansoni* populations fall close to the center of this range, with π being reduced by 51% relative to wild populations, a level of diversity reduction comparable to Mediterranean brown trout [29] and sunflower [30]. The relatively high levels of retained variation in laboratory schistosome populations may result from: (i) the relatively large size of *S. mansoni* founder populations; laboratory schistosome populations are typically founded by collecting eggs from one or more patients, each of which may be infected with hundreds of adult worms [31]; and (ii) that laboratory schistosome populations are maintained in quite large populations to prevent loss during laboratory maintenance.

Our census estimates show that numbers of independent schistosome genotypes used to infect hamsters ranges from 93-157 (harmonic means) in our laboratory populations.

The level of variation retained within populations is dependent on the size and duration of population bottlenecks as demonstrated with our population bottleneck simulation [32]. Our N_e estimates are 2 – 258 with NeEstimator and 5 – 123 with COLONY, while our census (N_c) estimates range from 93 to 157. The observed 51% reduction in nucleotide diversity (π) compared to field population variation is generally compatible with the simulation results when N ≈ 400. This is approximately double the observed N_c or N_e values estimated. However, we note that factors other than demographics may maintain genetic variation: both the action of balancing selection [33] or preferential mating between unrelated parasites [34] may also act to retain genetic variation in laboratory parasite populations.

While our study revealed moderate loss in nucleotide diversity in laboratory schistosome populations, there were dramatic differences in the *pattern* of variation in laboratory and field populations. The patterns observed are consistent with strong bottlenecks during establishment and maintenance *of S. mansoni* colonies. We observed (i) loss of rare alleles: this is reflected in the positive Tajima's D for four of the five laboratory populations, while the field populations show negative Tajima's D. This is furthermore confirmed by the allele frequency spectra, which show a deficit in rare alleles and more alleles at intermediate frequencies compared to field populations. Population contraction in the laboratory is the most likely cause of the allele frequency spectra observed as intermediate frequency alleles are more likely to survive bottlenecks [35]. (ii) Reduction in LD decay: with the notable exception of SmBRE, we observed 6 – 19-fold slower decay of LD with physical distance on chromosomes in laboratory populations compared with field populations. This reduction is expected given the increased

231 levels of sib-mating, genetic drift, and reduced total number of recombination events in small laboratory
232 populations.

233

234 **The exception: SmBRE is depauperate and shows low fitness**

235 We have studied the SmBRE laboratory population extensively. These parasites typically show reduced
236 snail infectivity, lower cercarial shedding and virulence in the intermediate snail host, and reduced
237 immunopathology in the mouse [36–39]. One possible explanation for low fitness of SmBRE is inbreeding
238 depression. In line with this, we found that nucleotide diversity (π) was two- to threefold lower in SmBRE
239 than in the other four laboratory populations; SmBRE also had the lowest estimates for effective
240 population size (N_e). However, other results were entirely unexpected. While SmEG, SmNMRI, SmLE, and
241 SmOR showed strongly positive Tajima's D, SmBRE had a strongly negative Tajima's D like the field
242 collected populations. We do not know what features of SmBRE demography might have contributed to
243 this.

244 LD analysis for SmBRE produced the most puzzling result and showed a pattern that was radically
245 different from the other laboratory populations. Given the low genetic diversity and effective population
246 size (N_e) in SmBRE, we had expected to see the slowest rate of LD decay among all groups in this
247 population. However, LD decayed very rapidly in SmBRE, and was higher than three of the four field
248 populations examined. A microsatellite based genetic map for *S. mansoni* revealed that 1cM = 227 kb
249 (95% CI 181 to 309kb) [41]. A possible explanation for the rapid decay of LD is that the recombination
250 rate may be higher in SmBRE than in other laboratory populations. Analysis of further *S. mansoni* genetic

crosses involving SmBRE could be used to explore this hypothesis. It is known that recombination rates can vary both across the genome and among populations of the several species [42,43].

Implication for schistosome research

That four of five laboratory-maintained schistosome populations retain abundant genetic variation has several important implications for schistosome research:

Phenotype measures in individual worms

Current research on schistosome parasites, including developmental, immunological, transcriptomic, or drug response studies, utilizes pools of genetically variable worms rather than homogeneous inbred parasite lines [44–47]. As a consequence, these studies capture average population phenotypes and underestimate variation in the traits studied. For example, our laboratory has recently documented a significant impact of parasite population on immunopathological parameters, including spleen/liver weight and fibrosis [36]. However, we recognize that these impacts are likely to be underestimated, as our studies, like many others, utilize genetically variable laboratory populations. Analysis of praziquantel response provides a dramatic example. The SmLE-PZQ-R laboratory population, selected for praziquantel resistance, shows 14-fold increase in drug resistance relative to SmLE population from which it was selected. However, SmLE-PZQ-R is a mixture of both PZQ sensitive (PQZ-S) and PZQ resistant (PQZ-R) parasites that differ by > 377 fold in drug response [17]. We suspect other parasite phenotypes may show equally dramatic variation when measured in individual worms rather than diverse populations. Open source tools like the Single Worm Analysis of Movement Pipeline (SWAMP) [48] and wrmXpress [49] now offer the capability to accurately measure drug response phenotypes in individual worms, while

transcriptomic variation can be measured in single worms or single cells [50–53]. We encourage researchers to shift their focus from genetically diverse populations to individual parasites for clearer measurement of parasite phenotypes.

Genome-wide association studies (GWAS)

Schistosome parasites show abundant phenotypic variation in a wide range of traits [54]. These include cercarial shedding [15,37,38,55–57], host specificity [58–60], and drug resistance [17,61–64]. Our laboratory is specifically interested in understanding the genetic basis of phenotypic traits in schistosomes, and we have primarily used genetic crosses and linkage analysis for this purpose [54]. That high levels of genetic variation are found within laboratory populations allows us to use a second powerful mapping approach (GWAS) to identify genes underlying specific traits. GWAS is considerably simpler than linkage analysis, because conducting two-generation (F2) genetic crosses is not required. Furthermore, GWAS more effectively examine variation across multiple individuals within populations, while genetic crosses examine differences between the two parents only, so samples genetic and phenotypic variation less effectively. Le Clec'h et al.'s [17] work on PZQ resistance provides strong proof-of-principal for use of GWAS approaches for schistosomes using laboratory populations. Their GWAS study used single worm measures of drug response in the SmLE-PZQ-R population and then sequenced pools of PZQ-S and PZQ-R worms showing extreme drug response phenotypes to determine the genome regions involved [17,65].

GWAS relies on association (LD) between trait loci and surrounding genetic markers. We observed much slower decay in LD in four out of five laboratory-maintained schistosome populations than observed in the field. GWAS studies in laboratory populations are therefore likely to generate much

293 broader peaks [66,67]. For example, in the GWAS of praziquantel resistance locus, the genome region
294 mapped spanned 5.72 mb and 137 genes [65]. Broad peaks have some advantages, as such peaks are
295 unlikely to be missed if they are situated in genome regions that are difficult to genotype. However,
296 broad peaks containing multiple genes make the task of identifying the causative locus much harder. We
297 note that the extremely rapid decay in LD observed in some field populations (e.g. Tanzania) suggests
298 that GWAS using freshly isolated parasite populations collected from infected patients may result in
299 narrow peaks and allow identification of candidate regions with greater precision.

300 Reproducibility at different institutions

301 Several laboratories maintain the same and/or different schistosome populations as examined here. The
302 literature often refers to these schistosome populations as strains or lines, akin to bacterial clones or
303 inbred mice, and so the assumption is that they will produce similar results at different institutions.
304 However, bottlenecks and low N_e will result in genetic drift, and divergence between populations at
305 different institutions. Such changes are likely to affect reproducibility, as is the case with non-model
306 rodents [68]. Accessing schistosome parasites through the BRI [3] increases short-term consistency and
307 reliability, but even parasites obtained from BRI in different years may vary due to genetic drift. While
308 genetically variable laboratory populations have advantages for some genetic analyses (e.g. see
309 “Genome-wide association studies”), one possible solution to increase repeatability in laboratory
310 experiments might be to establish inbred parasite lines by serial inbreeding over a minimum of seven
311 generations to reach 99% homozygosity. Such inbred lines have been used for snails [69] and mice [70]:
312 the addition of inbred schistosome lines would allow precise dissection of parasite host interactions
313 across the parasite life cycle. However, our experience with SmBRE illustrates that highly inbred

314 populations may suffer from inbreeding depression and reduced fitness, posing a significant challenge
315 to overcome.

316

317 **Relevance to other helminths**

318 Recent work suggests that the degree of genetic variation in other laboratory-maintained helminths
319 could be underestimated as well. Stevens et al. [33] showed that *Heligosomoides bakeri,* a commonly
320 used model nematode of rodents*,* retains extensive genetic diversity despite laboratory maintenance
321 for 70 years [33]. Even in the selfing nematode *C. elegans,* long-term balancing selection maintains
322 genetic variation to increase fitness and survival [71]. In line with these studies, sequencing of other
323 laboratory-maintained helminths, such as *Brugia pahangi* and *Trichuris muris*, may well provide similar
324 results. We anticipate that substantial genetic diversity will be found in these populations, therefore
325 providing new research opportunities for a wide range of model helminth parasites.

326

327 **Limitations of this study**

328 Paired field and laboratory populations from the same location are most informative for examining the
329 impact of laboratory culture or domestication. These were not available here, so we compared
330 laboratory sequence variation with that from previously published, but independent field collected
331 samples. We used Illumina short read sequencing for this work. Highly variable genes are difficult to align
332 to a reference sequence, so are typically excluded from illumina-based resequencing studies. It is
333 therefore likely that we significantly underestimated diversity in this study. For a more exhaustive

334 evaluation of genetic variation within laboratory schistosome populations, long read (Nanopore)
335 sequencing, Hi-C and *de novo* assembly will be needed [33]. For the same reason, our study was not well
336 powered to detect islands of genetic variation, suggestive of balancing selection, as observed in *C.*
337 *elegans* [71] and *H. bakeri* [33] (see “Relevance to other helminths”).

METHODS

Ethics statement

This study was performed in accordance with the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. The protocol was approved by the Institutional Animal Care and Use Committee of Texas Biomedical Research Institute (permit number: 1419-MA).

Recovery of *Schistosoma mansoni* miracidia and snail infections for sample generation

The experimental design used to generate our samples is summarized in Figure 2, and the methodology for each stage is explained below. We extracted gDNA from cercarial larvae in lieu of adult worms for two reasons: (i) adult schistosome females carry fertilized eggs which would result in mixed genotype sequences, and (ii) we wanted to avoid the sampling of identical adult worms derived from clonal cercariae from a single snail. In brief, we recovered *S. mansoni* eggs from livers of infected Golden Syrian hamsters as previously described [72] and infected *Biomphalaria glabrata* (line Bg36 for SmOR and SmLE) and *B. alexandrina* (for SmEG) snails by placing individuals in 24-well plates with a single miracidium. Plates were placed under a light source overnight before putting the snails in trays covered with a clear plastic lid. The lids were exchanged for a dark lid three weeks post infection to prevent cercarial shedding.

354 **Sample generation for SmBRE**

355 Preliminary analyses for this project revealed that our SmBRE population was already contaminated with
356 SmLE [19]. We therefore extracted gDNA from schistosome parasites previously collected during life
357 cycle maintenance. Individual male worms were processed as described below. To avoid obtaining mixed
358 genotype eggs, we decapitated individual female worms and extracted gDNA with Chelex® solution
359 following an established protocol [41]. All samples were whole-genome amplified as described below.

360

361 **Collection of *S. mansoni* cercariae and gDNA extraction**

362 We placed all snails into 24-well plates and shed them for 2 hours under light 28 days post infection. The
363 content in each individual well was collected, transferred into microtubes, and spun down at 500 × *g* for
364 5 minutes to pellet the cercariae. We removed supernatant before flash-freezing cercariae in liquid
365 nitrogen. Samples were stored at -80°C until gDNA extraction with the DNeasy Blood & Tissue Kit (Qiagen,
366 Germantown, MD, USA) according to manufacturer instructions (tissue lysis for 2 hours at 56°C). We
367 quantified gDNA using a Qubit dsDNA BR Assay Kit (Invitrogen, Carlsbad, CA, USA). We used the
368 GenomiPhi V2 DNA Amplification Kit for whole genome amplification (WGA) of samples with gDNA yield
369 < 200 ng (Cytiva, Marlborough, Massachusetts, USA).

370

371 **gDNA Library preparation and sequencing**

372 We used the KAPA HyperPlus Kit with library amplification (Roche, Indianapolis, IN, USA) to generate
373 whole genome libraries with 200-400 ng of input material. We followed the manufacturer's instructions

with the following modifications: we fragmented the samples for 25 minutes, amplified libraries using six PCR cycles, and we performed library size selection using a ratio of 0.6X (30 µl beads) for the first size cut and 0.8X (10 µl beads) for the second size cut. We assessed the library profile with TapeStation 4200 D1000 ScreenTape (Agilent, Santa Clara, CA, USA) (average library size: 455) and quantified all libraries with the KAPA Library Quantification Kit (Roche, Indianapolis, IN, USA) (average library concentration: 43 nM). Pooled samples were sent to Admera Health and sequenced on a NovaSeq S4 (one pool with 40 samples) or NovaSeq X Plus (3 pools with 18-19 samples) platform (Illumina) using 150 bp paired-end reads.

Computational environment

We used conda version 23.1.0 to manage environments and download packages used in the analysis. Data was processed in R 4.2.0 using *tidyverse* v1.3.2, and plots were generated with *ggplot* v3.4.2. All shell and R scripts written for this project are available at https://github.com/kathrinsjutzeler/sm_single_gt and Zenodo https://doi.org/10.5281/zenodo.10672479.

Genotyping

We used trim_galore v0.6.7 [73] (-q 28 --illumina --max_n 1 --clip_R1 9 --clip_R2 9) for adapter and quality trimming before mapping the sequences to version 9 of the *S. mansoni* reference genome (GenBank assembly accession GCA_000237925.5) with BWA v0.7.17-r118 [74] and the default

parameters. We used GATK v4.3.0.0 [75] for further processing of the sequences. First, we removed all optical/PCR duplicates with MarkDuplicates. Next, we called single nucleotide polymorphisms (SNPs) with HaplotypeCaller and GenotypeGVCFs on a contig-by-contig basis, which we combined for each individual and finally merged into a single VCF file for all sequences, including the ones from previously processed field samples [18]. At this point, we lifted the file over to v10 of the *S. mansoni* reference genome (Wellcome Sanger Institute, project PRJEA36577) using LiftoverVcf. We used VariantFiltration with the recommended parameters (FS > 60.0, SOR > 3.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, QD < 2.0) and VCFtools v0.1.16 [76] for quality filtering. For variant statistics of genomic data from laboratory populations, we removed sites with quality < 15, read depths < 10, and missingness > 20 % and individuals with a genotyping rate < 50%. For the combined laboratory/field population analyses of exome data, we i) removed sites with quality < 15, read depth < 10 and > 56% missingness, and ii) individuals with a genotyping rate < 50%. Finally, we used bedtools intersect (v2.31.0) [77] to keep only variants in the CDS region to normalize whole genome and exome data.

Principal component analysis (PCA) and admixture

We used the snpgdsPCA() function from the *SNPRelate* v1.30.1 [78] R package to generate the PCA matrix and ADMIXTURE v1.3.0 [79] to estimate population ancestry for which we examined between $k = 1$ and $k = 10$ populations. In the end, we chose the model with the smallest cross validation score and used Q estimates as a proxy for ancestry fractions.

Summary statistics, Tajima's D, and nucleotide diversity (π)

We calculated coverage statistics with samtools v1.9 [80] and mosdepth v0.3.6 [81]. We used VCFtools to calculate Tajima's D in windows of 25 kb using autosomal variants in each population separately. We generated a VCF file containing both variant and non-variant sites from the genotyped GATK database to calculate nucleotide diversity (π) in 25 kb windows with pixy [82].

Allele frequency spectrum and empirical cumulative distribution function (ECDF)

We used the site.spectrum() function from the *pegas v1.2* R package [83] to compute the folded site frequency spectrum and bcftools v1.9 [80] to get overall allele frequency for SNPs in each individual population. We used stat_ecdf() from *ggplot* to calculate and plot ECDF for a statistical comparison of laboratory and field populations.

Linkage disequilibrium

We examined linkage disequilibrium (LD) between autosomal variants within each population with PLINK v1.90b6.21 to make pairwise comparisons between SNPs within 1Mb of one another (--ld-window-r2 0.0, --ld-window 1000000, --ld-window-kb 1000). We binned average r^2 values using stats.bin() from the *fields* v14.1 R package [84] into 1,000 equal windows along the log scale which were calculated with logseq() from the *pracma v2.4.4* package [85]. Rare variants (MAF < 0.05) were excluded from this analysis. To compare LD decay curves, we measured the distance at which LD is reduced to $r^2 = 0.5$ ($LD_{0.5}$).

435 **Census and Ne estimation**

436 Census: We estimated census data using detailed schistosome life cycle maintenance records we keep
437 for each of our laboratory populations. Generally, we infect individual snails with five to ten miracidia
438 and record the number of infected and uninfected snails at the time of the first shedding. Therefore, the
439 probability of a snail not being infected is:

440 $$P(0) = \frac{Number\ of\ uninfected\ snails}{Number\ of\ surviving\ snails}$$

441 We then computed the probabilities of snail infections with varying numbers of miracidia utilizing a
442 Poisson distribution with the dpois() function from the *stats v4.2* package [86,87].

443 N_e estimation: We used two programs to determine effective population size: NeEstimator v2 [88], which
444 relies on linkage disequilibrium between pairs of SNPs on different chromosomes to estimate N_e and
445 COLONY v2 [89], which calculates N_e based on sibship inference. We used the R package *radiator* v1.2.8
446 [90] to convert working VCF files per population (14,073 – 119,643 loci) to suitable input files for each
447 software and ran COLONY via the command line with default parameters. Additionally, we created an
448 input file listing chromosomes and loci to run NeEstimator v2 with the "LD Locus Pairing" option which
449 excludes the comparison of loci on the same chromosome.

450

451 **Bottleneck simulation**

452 We used *vcfR* v1.13.0 [91] to extract genotypes from a VCF file containing common variants in the
453 Brazilian field population. We then randomly sampled 10,000 loci to generate an input file suitable for
454 BottleSim v2.6 [92]. We simulated bottleneck events with the "Diploid multilocus, constant population

455 size" option, assumed a generation overlap of 0 or 100, and dioecy with random mating. We ran this
456 simulation for N = 400, 200, 100, 50, 25, and 5 for 400 generations with a 1:1 sex ratio.

457

458 **Statistical analysis**

459 We performed all statistical analyses with R package *rstatix* v0.7.*2* [93] or *stats* v4.2. We used Student's
460 t-tests (parametric) or Wilcoxon's rank-sum test (non-parametric) to compare the means of field and
461 laboratory populations (normally distributed data, Shapiro test, $p > 0.05$). Comparisons between
462 empirical cumulative distributions were tested with the Kolmogorov-Smirnov test. We considered
463 comparisons statistically significant when $p < 0.05$ [32].

DECLARATIONS:

COMPETING INTERESTS:

The authors declare that they have no competing interests.

FUNDING:

This research was supported by a Graduate Research in Immunology Program training grant NIH T32 AI138944 (KSJ), and NIH R21 AI171601-02 (FDC, WL), R01 AI133749, R01 AI166049 (TJCA), and was conducted in facilities constructed with support from Research Facilities Improvement Program grant C06 RR013556 from the National Center for Research Resources. SNPRC research at Texas Biomedical Research Institute is supported by grant P51 OD011133 from the Office of Research Infrastructure Programs, NIH.

AUTHOR CONTRIBUTIONS:

KSJ and TJCA designed and planned the experiments. WL and FDC provided training, assisted with methodology, and provided long-term maintenance records. KSJ performed all the experiments (snail infections, collection of cercariae, DNA extraction, library preparation) and conducted the bioinformatics analysis with help from RNP who provided code and valuable guidance. RD and MM maintained the schistosome life cycle and recorded census data. KSJ and TJCA drafted the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS:

Snails infected with SmNMRI parasites were provided by the Schistosomiasis Resource Center of the Biomedical Research Institute (Rockville, MD) through NIH-NIAID Contract HHSN272201700014I. We thank Sarah Schmid and Gabrielle Bate for conducting the monomiracidial snail infections and coordinating shipping and Dr. Margaret Mentink-Kane for her assistance.

Data Availability Statement

The sequencing data generated for this project are available on Sequence Read Archive (SRA) under BioProject PRJNA1074697 (SmEG, SmOR, SmLE, SmNMRI) and PRJNA1170908 (SmBRE). Exome sequences from field samples have previously been published by Platt et al. [18] and are available on SRA under BioProjects PRJNA743359 (Brazil) and PRJNA560070 (Niger, Senegal, and Tanzania)

REFERENCES:

1. Stirewalt M, Cousin CE, Lewis FA, Leefe JL. Cryopreservation of Schistosomules of *Schistosoma Mansoni* in Quantity *. The American Journal of Tropical Medicine and Hygiene. 1984;33: 116–124. doi:10.4269/ajtmh.1984.33.116

2. Lewis FA, Stirewalt MA, Souza CP, Gazzinelli G. Large-scale laboratory maintenance of *Schistosoma mansoni*, with observations on three schistosome/snail host combinations. J Parasitol. 1986;72: 813–829.

3. Cody JJ, Ittiprasert W, Miller AN, Henein L, Mentink-Kane MM, Hsieh MH. The NIH-NIAID Schistosomiasis Resource Center at the Biomedical Research Institute: Molecular Redux. PLoS Negl Trop Dis. 2016;10: e0005022. doi:10.1371/journal.pntd.0005022

4. Hassan AHM, Haberl B, Hertel J, Haas W. Miracidia of an Egyptian Strain of *Schistosoma mansoni* Differentiate Between Sympatric Snail Species. Journal of Parasitology. 2003;89: 1248–1250. doi:10.1645/GE-85R

5. Botros SS, Hammam OA, El-Lakkany NM, El-Din SHS, Ebeid FA. Schistosoma haematobium (Egyptian strain): rate of development and effect of praziquantel treatment. J Parasitol. 2008;94: 386–394. doi:10.1645/GE-1270.1

6. Fneich S, Théron A, Cosseau C, Rognon A, Aliaga B, Buard J, et al. Epigenetic origin of adaptive phenotypic variants in the human blood fluke *Schistosoma mansoni*. Epigenetics & Chromatin. 2016;9: 27. doi:10.1186/s13072-016-0076-2

7. Rogers SH, Bueding E. Hycanthone resistance: development in Schistosoma mansoni. Science. 1971;172: 1057–1058. doi:10.1126/science.172.3987.1057

8. Fletcher M, LoVerde PT, Woodruff DS. Genetic variation in *Schistosoma mansoni*: enzyme polymorphisms in populations from Africa, Southwest Asia, South America, and the West Indies. Am J Trop Med Hyg. 1981;30: 406–421. doi:10.4269/ajtmh.1981.30.406

9. Minchella DJ, Lewis FA, Sollenberger KM, Williams JA. Genetic diversity of *Schistosoma mansoni*: quantifying strain heterogeneity using a polymorphic DNA element. Mol Biochem Parasitol. 1994;68: 307–313. doi:10.1016/0166-6851(94)90175-9

10. Pinto PM, Brito CF, Passos LK, Tendler M, Simpson AJ. Contrasting genomic variability between clones from field isolates and laboratory populations of *Schistosoma mansoni*. Mem Inst Oswaldo Cruz. 1997;92: 409–414. doi:10.1590/s0074-02761997000300019

11. Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, et al. The genome of the blood fluke *Schistosoma mansoni*. Nature. 2009;460: 352–358. doi:10.1038/nature08160

12. Protasio AV, Tsai IJ, Babbage A, Nichol S, Hunt M, Aslett MA, et al. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. PLoS Negl Trop Dis. 2012;6: e1455. doi:10.1371/journal.pntd.0001455

13. Protasio AV, Tsai IJ, Babbage A, Nichol S, Hunt M, Aslett MA, et al. A Systematically Improved High Quality Genome and Transcriptome of the Human Blood Fluke *Schistosoma mansoni*. Hoffmann KF, editor. PLoS Negl Trop Dis. 2012;6: e1455. doi:10.1371/journal.pntd.0001455

14. Davies CM, Webster JP, Woolhouse MEJ. Trade–offs in the evolution of virulence in an indirectly transmitted macroparasite. Proc R Soc Lond B. 2001;268: 251–257. doi:10.1098/rspb.2000.1367

15. Gower CM, Webster JP. Fitness of indirectly transmitted pathogens: restraint and constraint. Evolution. 2004;58: 1178–1184. doi:10.1111/j.0014-3820.2004.tb01698.x

16. Couto FFB, Coelho PMZ, Araújo N, Kusel JR, Katz N, Jannotti-Passos LK, et al. *Schistosoma mansoni*: a method for inducing resistance to praziquantel using infected *Biomphalaria glabrata* snails. Mem Inst Oswaldo Cruz. 2011;106: 153–157. doi:10.1590/s0074-02762011000200006

17. Le Clec'h W, Chevalier FD, Mattos ACA, Strickland A, Diaz R, McDew-White M, et al. Genetic analysis of praziquantel response in schistosome parasites implicates a transient receptor potential channel. Sci Transl Med. 2021;13: eabj9114. doi:10.1126/scitranslmed.abj9114

18. Platt RN, Le Clec'h W, Chevalier FD, McDew-White M, LoVerde PT, de Assis RR, et al. Genomic analysis of a parasite invasion: colonization of the Americas by the blood fluke, *Schistosoma mansoni*. Evolutionary Biology; 2021 Oct. doi:10.1101/2021.10.25.465783

19. Jutzeler KS, Platt RN, Li X, Morales M, Diaz R, Clec'h WL, et al. Rapid phenotypic and genotypic change in a laboratory schistosome population. bioRxiv. 2024; 2024.08.06.606850. doi:10.1101/2024.08.06.606850

20. Dutrow EV, Serpell JA, Ostrander EA. Domestic dog lineages reveal genetic drivers of behavioral diversification. Cell. 2022;185: 4737-4755.e18. doi:10.1016/j.cell.2022.11.003

21. Bower MA, McGivney BA, Campana MG, Gu J, Andersson LS, Barrett E, et al. The genetic origin and history of speed in the Thoroughbred racehorse. Nat Commun. 2012;3: 643. doi:10.1038/ncomms1644

22. Brotherstone S, Goddard M. Artificial selection and maintenance of genetic variance in the global dairy cow population. Philos Trans R Soc Lond B Biol Sci. 2005;360: 1479–1488. doi:10.1098/rstb.2005.1668

23. Cole JB, Makanjuola BO, Rochus CM, van Staaveren N, Baes C. The effects of breeding and selection on lactation in dairy cattle. Anim Front. 2023;13: 55–63. doi:10.1093/af/vfad044

24. Núñez-León D, Cordero GA, Schlindwein X, Jensen P, Stoeckli E, Sánchez-Villagra MR, et al. Shifts in growth, but not differentiation, foreshadow the formation of exaggerated forms under chicken domestication. Proc Biol Sci. 2021;288: 20210392. doi:10.1098/rspb.2021.0392

25. Liu W, Chen L, Zhang S, Hu F, Wang Z, Lyu J, et al. Decrease of gene expression diversity during domestication of animals and plants. BMC Evol Biol. 2019;19: 19. doi:10.1186/s12862-018-1340-9

26. Albert FW, Somel M, Carneiro M, Aximu-Petri A, Halbwax M, Thalmann O, et al. A comparison of brain gene expression levels in domesticated and wild animals. PLoS Genet. 2012;8: e1002962. doi:10.1371/journal.pgen.1002962

27. Hu C, Yuan S, Sun W, Chen W, Liu W, Li P, et al. Spatial Genetic Structure and Demographic History of the Wild Boar in the Qinling Mountains, China. Animals (Basel). 2021;11: 346. doi:10.3390/ani11020346

28. Zhang J, Yang B, Wen X, Sun G. Genetic variation and relationships in the mitochondrial DNA D-loop region of Qinghai indigenous and commercial pig breeds. Cell Mol Biol Lett. 2018;23: 31. doi:10.1186/s11658-018-0097-x

29. Leitwein M, Gagnaire P-A, Desmarais E, Guendouz S, Rohmer M, Berrebi P, et al. Genome-wide nucleotide diversity of hatchery-reared Atlantic and Mediterranean strains of brown trout Salmo trutta compared to wild Mediterranean populations. J Fish Biol. 2016;89: 2717–2734. doi:10.1111/jfb.13131

30. Liu A, Burke JM. Patterns of nucleotide diversity in wild and cultivated sunflower. Genetics. 2006;173: 321–330. doi:10.1534/genetics.105.051110

31. Cheever AW, Kamel IA, Elwi AM, Mosimann JE, Danner R. *Schistosoma mansoni* and *S. haematobium* infections in Egypt. II. Quantitative parasitological findings at necropsy. Am J Trop Med Hyg. 1977;26: 702–716. doi:10.4269/ajtmh.1977.26.702

32. Nei M, Maruyama T, Chakraborty R. THE BOTTLENECK EFFECT AND GENETIC VARIABILITY IN POPULATIONS. Evolution. 1975;29: 1–10. doi:10.1111/j.1558-5646.1975.tb00807.x

33. Stevens L, Martínez-Ugalde I, King E, Wagah M, Absolon D, Bancroft R, et al. Ancient diversity in host-parasite interaction genes in a model parasitic nematode. Nat Commun. 2023;14: 7776. doi:10.1038/s41467-023-43556-w

34. Beltran S, Cézilly F, Boissier J. Genetic dissimilarity between mates, but not male heterozygosity, influences divorce in schistosomes. PLoS One. 2008;3: e3328. doi:10.1371/journal.pone.0003328

35. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, et al. Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res. 2005;15: 1553–1565. doi:10.1101/gr.4326505

36. Jutzeler KS, Le Clec'h W, Chevalier FD, Anderson TJC. Contribution of parasite and host genotype to immunopathology of schistosome infections. Parasit Vectors. 2024;17: 203. doi:10.1186/s13071-024-06286-6

37. Le Clec'h W, Chevalier FD, McDew-White M, Menon V, Arya G-A, Anderson TJC. Genetic architecture of transmission stage production and virulence in schistosome parasites. Virulence. 2021;12: 1508–1526. doi:10.1080/21505594.2021.1932183

38. Le Clec'h W, Diaz R, Chevalier F, McDew-White M, Anderson T. Striking differences in virulence, transmission and sporocyst growth dynamics between two schistosome populations. Parasites & Vectors. 2019;12: 485.

39. Le Clec'h W, Chevalier FD, Jutzeler K, Anderson TJC. No evidence for schistosome parasite fitness trade-offs in the intermediate and definitive host. Parasites Vectors. 2023;16: 132. doi:10.1186/s13071-023-05730-3

40. Campitelli BE, Stinchcombe JR. Population dynamics and evolutionary history of the weedy vine Ipomoea hederacea in North America. G3 (Bethesda). 2014;4: 1407–1416. doi:10.1534/g3.114.011700

41. Criscione CD, Valentim CLL, Hirai H, LoVerde PT, Anderson TJC. Genomic linkage map of the human blood fluke Schistosoma mansoni. Genome Biol. 2009;10: R71. doi:10.1186/gb-2009-10-6-r71

42. Dutheil JY. On the estimation of genome-average recombination rates. Genetics. 2024;227: iyae051. doi:10.1093/genetics/iyae051

43. Venu V, Harjunmaa E, Dreau A, Brady S, Absher D, Kingsley DM, et al. Fine-scale contemporary recombination variation and its fitness consequences in adaptively diverging stickleback fish. Nat Ecol Evol. 2024;8: 1337–1352. doi:10.1038/s41559-024-02434-4

44. Chalmers IW, McArdle AJ, Coulson RM, Wagner MA, Schmid R, Hirai H, et al. Developmentally regulated expression, alternative splicing and distinct sub-groupings in members of the *Schistosoma mansoni* venom allergen-like (SmVAL) gene family. BMC Genomics. 2008;9: 89. doi:10.1186/1471-2164-9-89

45. Kalantari P, Shecter I, Hopkins J, Pilotta Gois A, Morales Y, Harandi BF, et al. The balance between gasdermin D and STING signaling shapes the severity of schistosome immunopathology. Proc Natl Acad Sci U S A. 2023;120: e2211047120. doi:10.1073/pnas.2211047120

46. Lu Z, Sankaranarayanan G, Rawlinson KA, Offord V, Brindley PJ, Berriman M, et al. The Transcriptome of *Schistosoma mansoni* Developing Eggs Reveals Key Mediators in Pathogenesis and Life Cycle Propagation. Front Trop Dis. 2021;2: 713123. doi:10.3389/fitd.2021.713123

47. Mukendi JPK, Nakamura R, Uematsu S, Hamano S. Interleukin (IL)-33 is dispensable for *Schistosoma mansoni* worm maturation and the maintenance of egg-induced pathology in intestines of infected mice. Parasites Vectors. 2021;14: 70. doi:10.1186/s13071-020-04561-w

48. Chevalier FD. SWAMP: Single Worm Analysis of Movement Pipeline. Available: https://github.com/fdchevalier/SWAMP

49. Wheeler NJ, Gallo KJ, Rehborg EJG, Ryan KT, Chan JD, Zamanian M. wrmXpress: A modular package for high-throughput image analysis of parasitic and free-living worms. PLoS Negl Trop Dis. 2022;16: e0010937. doi:10.1371/journal.pntd.0010937

50. Diaz Soria CL, Lee J, Chong T, Coghlan A, Tracey A, Young MD, et al. Single-cell atlas of the first intra-mammalian developmental stage of the human parasite *Schistosoma mansoni*. Nat Commun. 2020;11: 6411. doi:10.1038/s41467-020-20092-5

51. Nanes Sarfati D, Li P, Tarashansky AJ, Wang B. Single-cell deconstruction of stem-cell-driven schistosome development. Trends Parasitol. 2021;37: 790–802. doi:10.1016/j.pt.2021.03.005

52. Wendt G, Zhao L, Chen R, Liu C, O'Donoghue AJ, Caffrey CR, et al. A single-cell RNA-seq atlas of *Schistosoma mansoni* identifies a key regulator of blood feeding. Science. 2020;369: 1644–1649. doi:10.1126/science.abb7709

53. Wendt GR, Reese ML, Collins JJ. SchistoCyte Atlas: A Single-Cell Transcriptome Resource for Adult Schistosomes. Trends Parasitol. 2021;37: 585–587. doi:10.1016/j.pt.2021.04.010

54. Anderson TJC, LoVerde PT, Le Clec'h W, Chevalier FD. Genetic Crosses and Linkage Mapping in Schistosome Parasites. Trends Parasitol. 2018;34: 982–996. doi:10.1016/j.pt.2018.08.001

55. Webster JP, Gower CM, Blair L. Do hosts and parasites coevolve? Empirical support from the *Schistosoma* system. Am Nat. 2004;164 Suppl 5: S33-51. doi:10.1086/424607

56. Webster JP, Davies CM. Coevolution and compatibility in the snail-schistosome system. Parasitology. 2001;123 Suppl: S41-56. doi:10.1017/s0031182001008071

57. Théron A. Chronobiology of trematode cercarial emergence: from data recovery to epidemiological, ecological and evolutionary implications. Adv Parasitol. 2015;88: 123–164. doi:10.1016/bs.apar.2015.02.003

58. Mitta G, Gourbal B, Grunau C, Knight M, Bridger JM, Théron A. The Compatibility Between *Biomphalaria glabrata* Snails and *Schistosoma mansoni*: An Increasingly Complex Puzzle. Adv Parasitol. 2017;97: 111–145. doi:10.1016/bs.apar.2016.08.006

59. Rollinson D, Stothard JR, Southgate VR. Interactions between intermediate snail hosts of the genus *Bulinus* and schistosomes of the *Schistosoma haematobium* group. Parasitology. 2001;123 Suppl: S245-260. doi:10.1017/s0031182001008046

60. Theron A, Rognon A, Gourbal B, Mitta G. Multi-parasite host susceptibility and multi-host parasite infectivity: a new approach of the *Biomphalaria glabrata*/*Schistosoma mansoni* compatibility polymorphism. Infect Genet Evol. 2014;26: 80–88. doi:10.1016/j.meegid.2014.04.025

61. Valentim CLL, Cioli D, Chevalier FD, Cao X, Taylor AB, Holloway SP, et al. Genetic and Molecular Basis of Drug Resistance and Species-Specific Drug Action in Schistosome Parasites. Science. 2013;342: 1385–1389. doi:10.1126/science.1243106

62. Greenberg RM. New approaches for understanding mechanisms of drug resistance in schistosomes. Parasitology. 2013;140: 1534–1546. doi:10.1017/S0031182013000231

63. Melman SD, Steinauer ML, Cunningham C, Kubatko LS, Mwangi IN, Wynn NB, et al. Reduced susceptibility to praziquantel among naturally occurring Kenyan isolates of *Schistosoma mansoni*. PLoS Negl Trop Dis. 2009;3: e504. doi:10.1371/journal.pntd.0000504

64. Mwangi IN, Sanchez MC, Mkoji GM, Agola LE, Runo SM, Cupit PM, et al. Praziquantel sensitivity of Kenyan *Schistosoma mansoni* isolates and the generation of a laboratory strain with reduced susceptibility to the drug. Int J Parasitol Drugs Drug Resist. 2014;4: 296–300. doi:10.1016/j.ijpddr.2014.09.006

65. Chevalier FD, Le Clec'h W, Berriman M, Anderson TJC. A single locus determines praziquantel response in *Schistosoma mansoni*. Antimicrob Agents Chemother. 2024; e0143223. doi:10.1128/aac.01432-23

66. Christoforou A, Dondrup M, Mattingsdal M, Mattheisen M, Giddaluru S, Nöthen MM, et al. Linkage-disequilibrium-based binning affects the interpretation of GWASs. Am J Hum Genet. 2012;90: 727–733. doi:10.1016/j.ajhg.2012.02.025

67. Joiret M, Mahachie John JM, Gusareva ES, Van Steen K. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. BioData Min. 2019;12: 11. doi:10.1186/s13040-019-0199-7

68. Brekke TD, Steele KA, Mulley JF. Inbred or Outbred? Genetic Diversity in Laboratory Rodent Colonies. G3 (Bethesda). 2018;8: 679–686. doi:10.1534/g3.117.300495

69. Mulvey M, Woodruff DS, Carpenter MP. Linkage relationships of seven enzyme and two pigmentation loci in the snail Biomphalaria glabrata. J Hered. 1988;79: 473–476. doi:10.1093/oxfordjournals.jhered.a110554

70. Casellas J. Inbred mouse strains and genetic stability: a review. Animal. 2011;5: 1–7. doi:10.1017/S1751731110001667

71. Lee D, Zdraljevic S, Stevens L, Wang Y, Tanny RE, Crombie TA, et al. Balancing selection maintains hyper-divergent haplotypes in *Caenorhabditis elegans*. Nat Ecol Evol. 2021;5: 794–807. doi:10.1038/s41559-021-01435-x

72. Tucker MS, Karunaratne LB, Lewis FA, Freitas TC, Liang Y. Schistosomiasis. Current Protocols in Immunology. 2013;103. doi:10.1002/0471142735.im1901s103

73. Krueger F, James F, Ewels P, Afyounian E, Weinstein M, Schuster-Boeckler B. TrimGalore. Available: https://github.com/FelixKrueger/TrimGalore

74. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

75. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303. doi:10.1101/gr.107524.110

76. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330

77. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

78. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28: 3326–3328. doi:10.1093/bioinformatics/bts606

79. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664. doi:10.1101/gr.094052.109

80. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10: giab008. doi:10.1093/gigascience/giab008

81. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018;34: 867–868. doi:10.1093/bioinformatics/btx699

82. Korunes KL, Samuk K. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. Mol Ecol Resour. 2021;21: 1359–1368. doi:10.1111/1755-0998.13326

704 83. Paradis E. pegas: an R package for population genetics with an integrated–modular approach.
705 Bioinformatics. 2010;26: 419–420. doi:10.1093/bioinformatics/btp696

706 84. Douglas Nychka, Reinhard Furrer, John Paige, Stephan Sain. fields: Tools for spatial data. Boulder, CO, USA:
707 University Corporation for Atmospheric Research; 2021. Available:
708 https://github.com/dnychka/fieldsRPackage

709 85. Borchers HW. pracma: Practical Numerical Math Functions. 2023. Available: https://CRAN.R-
710 project.org/package=pracma

711 86. Gourbière S, Morand S, Waxman D. Fundamental factors determining the nature of parasite aggregation in
712 hosts. PLoS One. 2015;10: e0116893. doi:10.1371/journal.pone.0116893

713 87. McVinish R, Lester RJG. Measuring aggregation in parasite populations. J R Soc Interface. 2020;17:
714 20190886. doi:10.1098/rsif.2019.0886

715 88. Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR. NeEstimator v2: re-implementation of
716 software for the estimation of contemporary effective population size (N_e) from genetic data. Molecular
717 Ecology Resources. 2014;14: 209–214. doi:10.1111/1755-0998.12157

718 89. Jones OR, Wang J. COLONY: a program for parentage and sibship inference from multilocus genotype data.
719 Molecular Ecology Resources. 2010;10: 551–555. doi:10.1111/j.1755-0998.2009.02787.x

720 90. Gosselin T. thierrygosselin/radiator: update. Zenodo; 2020. doi:10.5281/ZENODO.3687060

721 91. Knaus BJ, Grünwald NJ. vcfr: a package to manipulate and visualize variant call format data in R. Mol Ecol
722 Resour. 2017;17: 44–53. doi:10.1111/1755-0998.12549

723 92. Kuo C -H., Janzen FJ. BOTTLESIM : a bottleneck simulation program for long-lived species with overlapping
724 generations. Molecular Ecology Notes. 2003;3: 669–673. doi:10.1046/j.1471-8286.2003.00532.x

725 93. Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. 2023. Available:
726 <https://CRAN.R-project.org/package=rstatix>

727

728

729 **FIGURE LEGENDS:**

730 **Figure 1: Genetic consequences of long-term laboratory maintenance.** Genetic variation in laboratory
731 schistosome populations may be removed or retained during repeated passage during life cycle
732 maintenance. We aimed to directly quantify levels of variation in laboratory *S. mansoni* populations.

733 **Figure 2: Sample generation:** We infected 192 – 240 *Biomphalaria glabrata* (SmOR and SmLE) or *B.*
734 *alexandrina* (SmEG) snails with a single *Schistosoma mansoni* miracidium. We shed the snails 28 days
735 post parasite exposure to identify infected (i.e. shedding) snails and to collect cercariae for gDNA library
736 preparation and sequencing. We used the bioinformatics pipeline outlined to analyze all the data. We
737 used adult worms from previous life cycle maintenance to generate sequences for SmBRE, as this
738 population was contaminated at the time of this experiment.

739 **Figure 3: Population structure in *S. mansoni* laboratory populations.** Both plots demonstrate the
740 separation of each population with the exception of SmBRE and SmLE. **(A)** PCA plot showing clustering
741 of sequenced *S. mansoni* laboratory populations. **(B)** Admixture analysis with $k = 5$ populations.

742 **Figure 4: Comparable nucleotide diversity in field and laboratory populations. (A)** Average nucleotide
743 diversity across the whole genome for each laboratory population calculated in 25 kb windows and
744 plotted for each autosome. The line indicates a LOESS smoothed curve. **(B)** Box and whisker plot showing
745 nucleotide diversity (π) in 25 kb windows across the CDS in laboratory and field populations. Outliers are
746 not shown.

747 **Figure 5: Indicators of recent bottlenecks in laboratory populations. (A)** Bar plots showing mean and
748 standard error of Tajima's D in each population. A *t*-test was used to compare means of Tajima's D in
749 field and laboratory populations. **(B)** Line plot showing the empirical cumulative distribution function

(ECDF) of allele frequencies in each population. Kolmogorov-Smirnov test was used to compare field vs laboratory distributions.

Figure 6: Slower LD decay in laboratory populations. (A) r^2 showing LD decay with physical distance between all autosomal SNPs in laboratory populations and exonic SNPs in field populations along the chromosomes. Mean was calculated over 1 kb windows following the log scale. **(B)** Bar plot showing position when $r^2 = 0.5$ ($LD_{0.5}$) for field and laboratory populations. A *t*-test was used to compare field and laboratory populations.

Figure 7: Reduced effective population size in laboratory populations. Bar plots showing effective population size N_e calculated with **(A)** NeEstimator and **(B)** COLONY. The y-axis is split to show both high and low N_e values clearly. The error bars represent a 95% confidence interval.

Figure 8: Bottleneck simulation over 400 generations with and without overlap. Line plot showing simulated reduction in genetic diversity of schistosome populations of different sizes over 400 generations. We used constant N ranging from 5 – 400. The horizontal dashed line shows 49% indicative of the retention of diversity observed in our laboratory populations.

Figure S1: Folded allele frequency spectra. Histograms of folded allele frequency spectra of each *S. mansoni* population.

Figure S2: Estimated census size (N_c) of laboratory *S. mansoni* populations. (A) Line plot showing estimated census size over time. We used detailed life cycle maintenance records to estimate P(0) and calculated numbers of parasites/snail assuming a Poisson distribution. Note that these N_c values are likely to be systematic overestimates. We conduct hamster infections with newly infected batches of snails to which we add surviving infected snails from the prior life cycle maintenance. Therefore, the

771 proportion of uninfected snails (P(0)) will be underestimated, and Poisson estimates of numbers of
772 parasite genotypes per snail will be overestimated. The actual N_c values are likely to be somewhat lower.
773 **(B)** Bar plot showing the harmonic mean of the N_c for each population. The error bars represent a 95%
774 confidence interval. **(C)** Scatter plot showing the relationship between N_e as calculated by COLONY (filled
775 circle) and NeEstimator (open circle) for each population. The lines represent a linear regression model,
776 and the corresponding Pearson correlation coefficients are displayed in accordance with the legend of
777 the tool used.

778 **Figure S3: LD decay between exonic SNPs in all *S. mansoni* populations. (A)** r^2 showing LD decay with
779 physical distance between exonic SNPs along the chromosomes. Mean was calculated over 1 kb windows
780 following the log scale except for SmBRE for which all data points were plotted. **(B)** Bar plot showing
781 position when $r^2 = 0.5$ ($LD_{0.5}$) for field and laboratory populations. A *t*-test was used to compare field and
782 laboratory populations.

Table 1 - Summary Statistics of Laboratory Populations

Population	Number of samples	Mean coverage (Range coverage)	All variants	SNVs	INDELS[1]	Autosomal SNPs	Mitochondrial SNPs	SNPs MAF > 0.05
BRE	20	71.1 (47.8, 143.8)	8.97E+05	8.11E+05	8.55E+04	7.37E+05	7	1.26E+05
EG	24	24.8 (17.3, 38.3)	1.22E+06	1.11E+06	1.10E+05	1.03E+06	9	8.69E+05
LE	24	23.4 (10.5, 44.5)	1.01E+06	9.15E+05	9.65E+04	8.62E+05	7	5.23E+05
NMRI	19	26.3 (15.9, 38.5)	1.08E+06	9.83E+05	9.35E+04	9.36E+05	2	7.23E+05
OR	21	24.4 (10.0, 42.2)	1.07E+06	9.55E+05	1.19E+05	9.23E+05	5	6.40E+05

[1] Mean INDEL size = -98, range (-369, 406)

Table 1 – continued

Population	Number of samples	Synonymous coding	Non-synonymous coding	Intron	Intergenic
BRE	20	9.65E+03	1.33E+04	4.64E+05	4.27E+05
EG	24	1.30E+04	1.53E+04	6.19E+05	5.95E+05
LE	24	1.03E+04	1.26E+04	5.12E+05	4.96E+05
NMRI	19	1.08E+04	1.33E+04	5.57E+05	5.20E+05
OR	21	1.10E+04	1.33E+04	5.51E+05	5.20E+05

[1] Mean INDEL size = -98, range (-369, 406)

Table 2 - Summary of variants used for the analyses

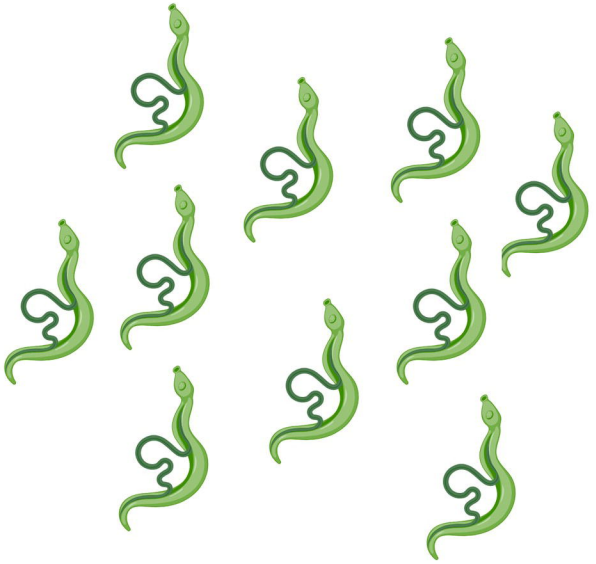
Population	Number of samples	Autosomal SNPs in CDS region	MAF filtered (> 0.05)
BRE	20	14,073	1,215
EG	24	18,574	14,573
LE	24	14,977	8,420
NMRI	19	15,504	11,581
OR	20	16,160	10,591
Brazil	43	41,599	21,691
Niger	9	31,710	29,539
Senegal	24	51,769	15,076
Tanzania	45	119,643	40,748

Table 3 – N_c, N_e estimates and ratios

Population	Census	Census CI95 (L)	Census CI95 (U)	NeEstimator	NeEstimator CI95 (L)	NeEstimator CI95 (U)	COLONY	COLONY CI95 (L)	COLONY CI95 (U)	NeEstimator N_e/N_c	COLONY N_e/N_c
BRE	93	79	114	2	2	2	5	2	20	0.02	0.05
EG	132	111	161	81	80	81	55	33	112	0.61	0.42
LE	157	127	205	258	253	264	123	71	417	1.65	0.79
OR	137	112	174	42	42	42	45	27	99	0.31	0.33
NMRI				237	232	242	114	60	581		
Brazil				3,174	3,064	3,291	3,612	1,211	Infinite		
Niger				Infinite	Infinite	Infinite	Infinite	1	Infinite		
Senegal				Infinite	Infinite	Infinite	Infinite	1	Infinite		
Tanzania				Infinite	Infinite	Infinite	Infinite	1	Infinite		

Table 4. Nucleotide diversity in other species populations

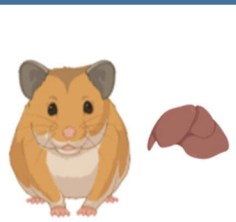
Organism	Wild	Domesticated/Farmed/ Laboratory adapted	Reduction in diversity	Authors
Animals				
Wolf/Dog	0.02-0.011	0.001-0.0004	95% (91- 98%)	Djan et al. (2014), Brouillette et al. (2000)
Boar/Pig	0.0079	0.00264 - 0.01559	-15% (-97 - 66%)	Hu et al. (2021), Zhang et al. (2018)
Chicken	0.0002	0.00010	37%	Zhang et al. (2018), Zhang et al. (2023)
Elliot's Pheasant	0.0063	0.0015	76%	Jiang et al. (2005)
Salmon	0.1114	0.04446	60%	Tsaparis et al. (2022)
Italian brown trout	0.0011	0.00043	61%	Magris et al. (2022)
Mediterranean brown trout	0.0049	0.0029-0.004	56% (49 - 63%)	Leitwein et al. (2016)
Drosophila	0.1557	0.0048	97%	Lian et al. (2017), Kapun et al. (2021)
Plants				
Teosinte	0.0097	0.0064	34%	Wright et al. (2005)
Alfalfa	0.0202	0.0135	33%	Muller et al (2006)
Sunflower	0.0128	0.0056	56%	Liu and Burke (2006)
Wheat	0.0023	0.0008	65%	Haudry et al. (2007)
Elephant Foot Yam	0.3058	0.08594	72%	Gao et al. (2017)
Schistosomes				
Schistosoma mansoni	**0.0014**	**0.0007**	**51%**	**This study**



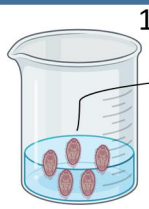
Genetically **homogeneous** – variation <u>removed</u> by serial inbreeding during laboratory passage.

Genetically **heterogeneous** – variation <u>retained</u> during laboratory passage.

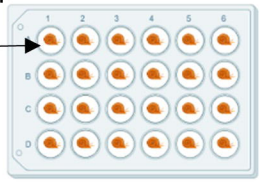
Sample generation



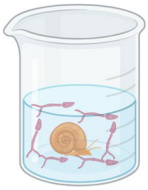
Liver from *S. mansoni* infected hamster



S. mansoni miracidia



192 – 240 snails per *S. mansoni* population



Infected snail shedding cercariae

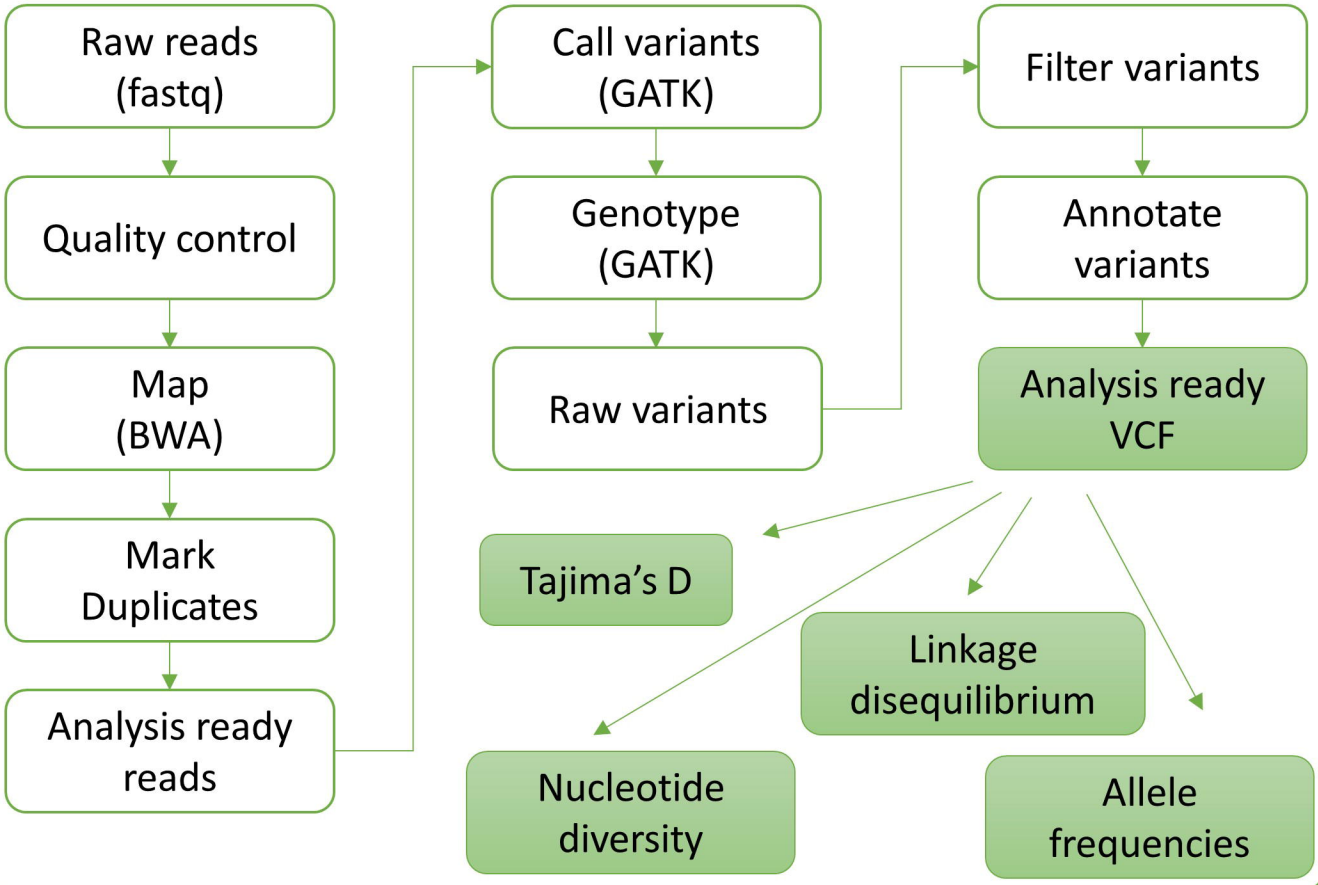
Library preparation and whole genome sequencing

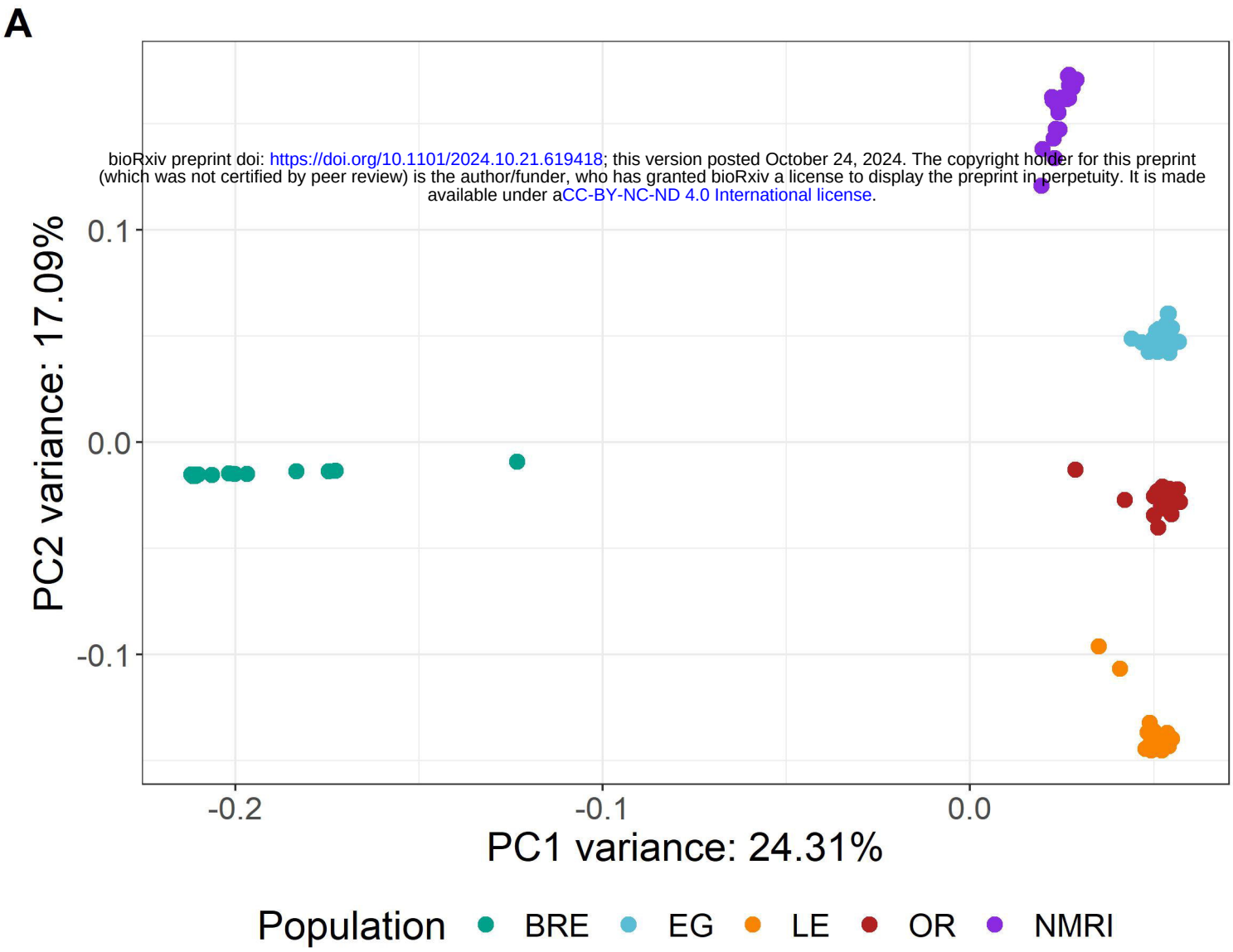
S. mansoni cercariae

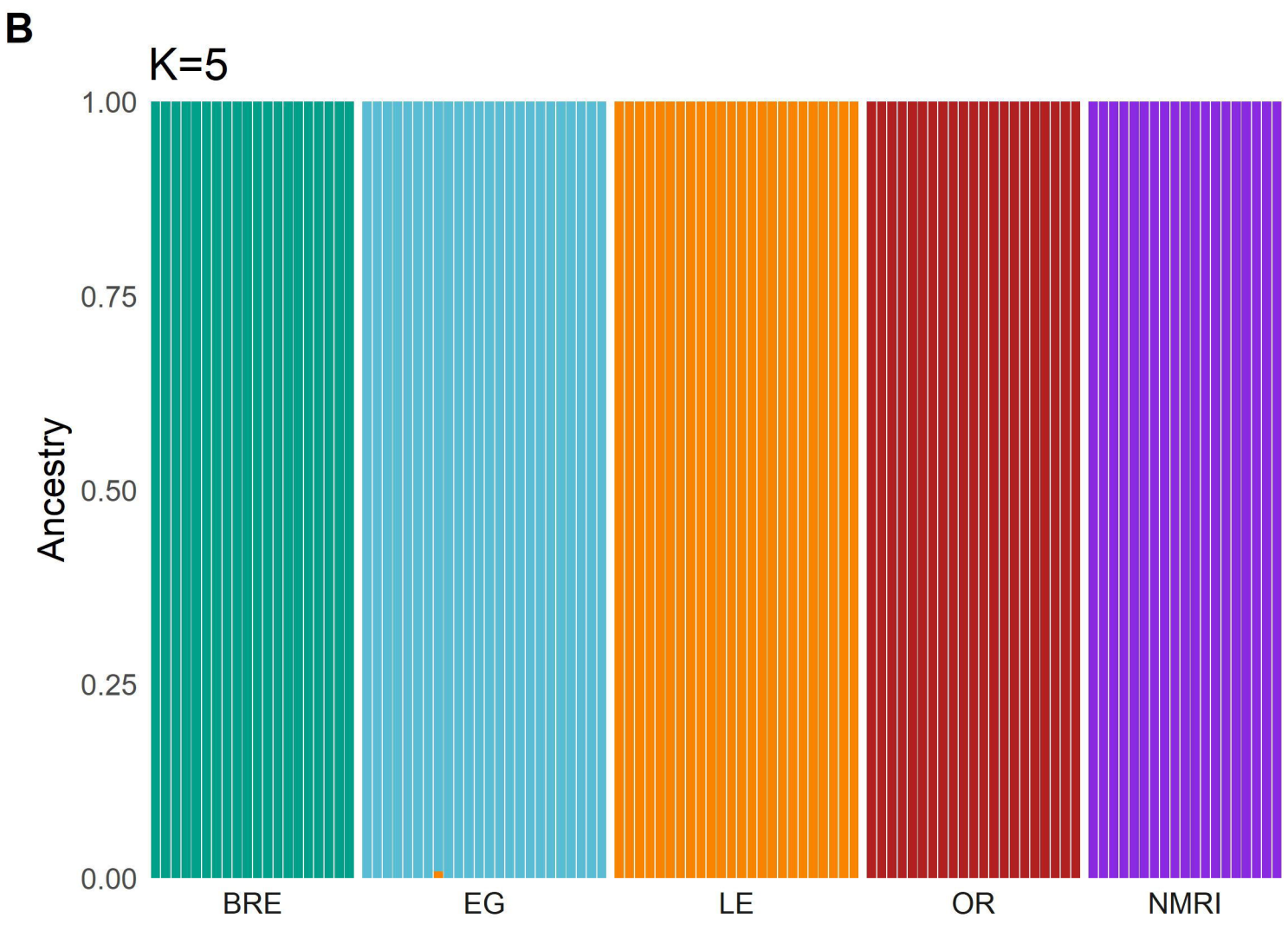
gDNA extraction

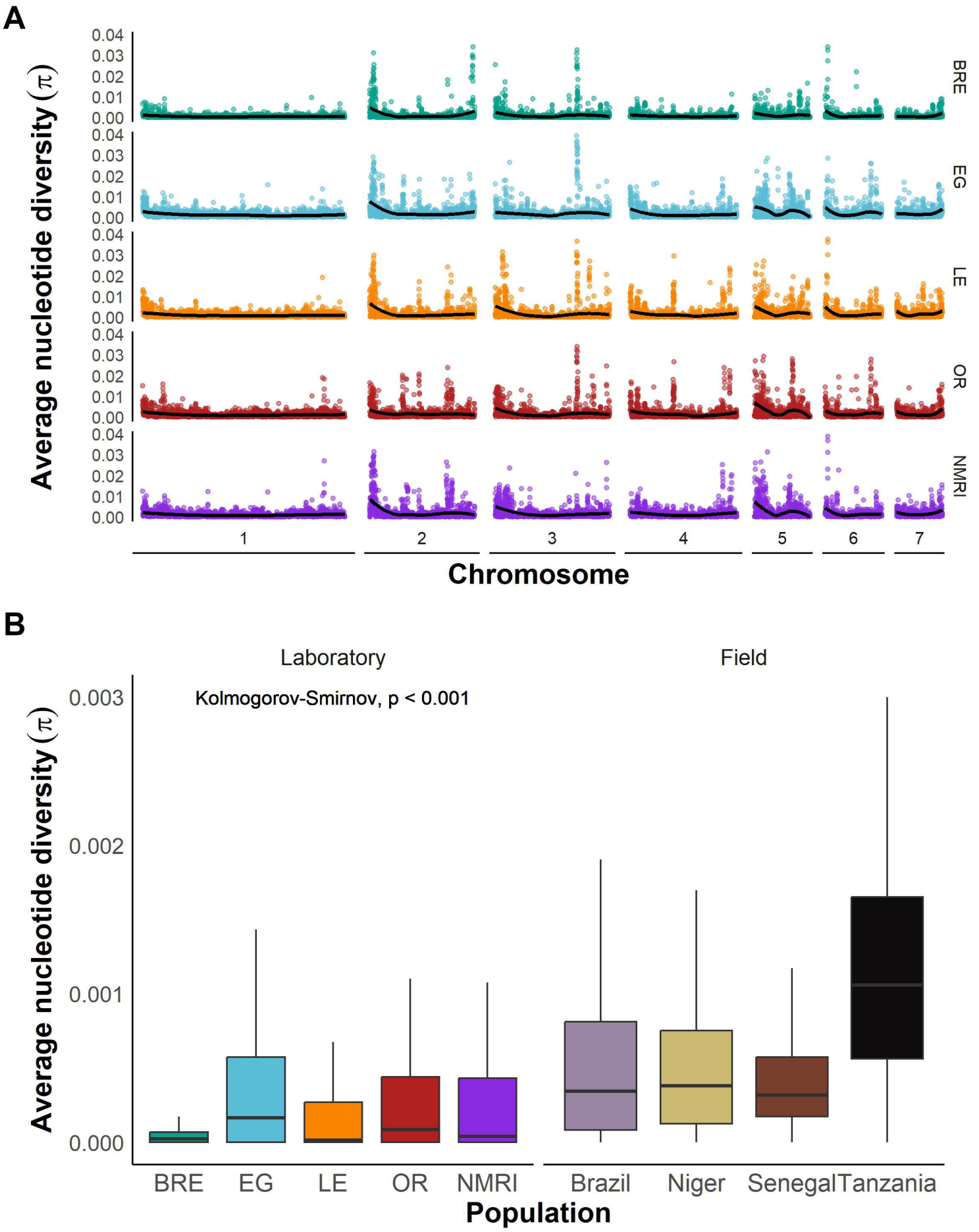
Whole genome library prep

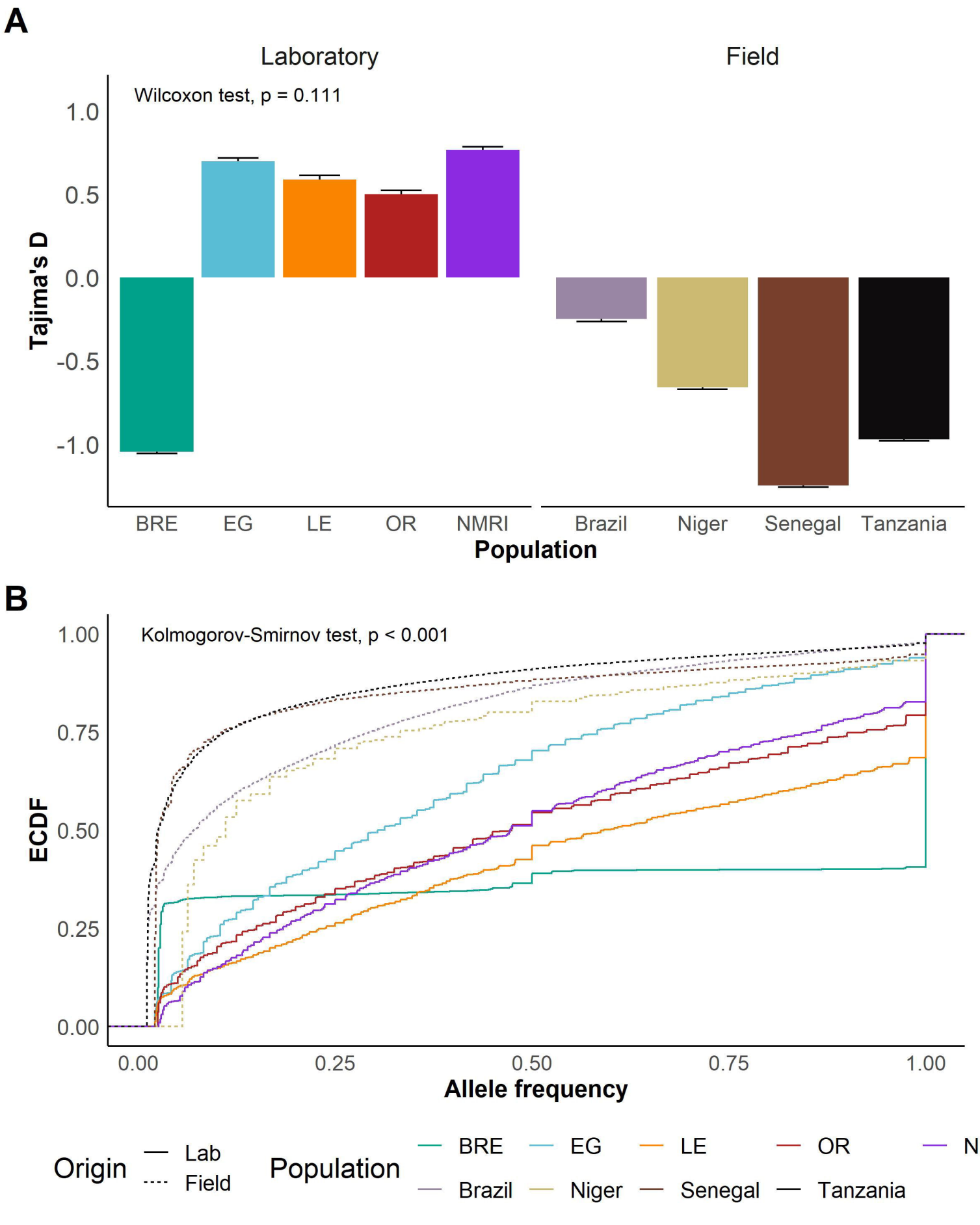
Bioinformatic pipeline

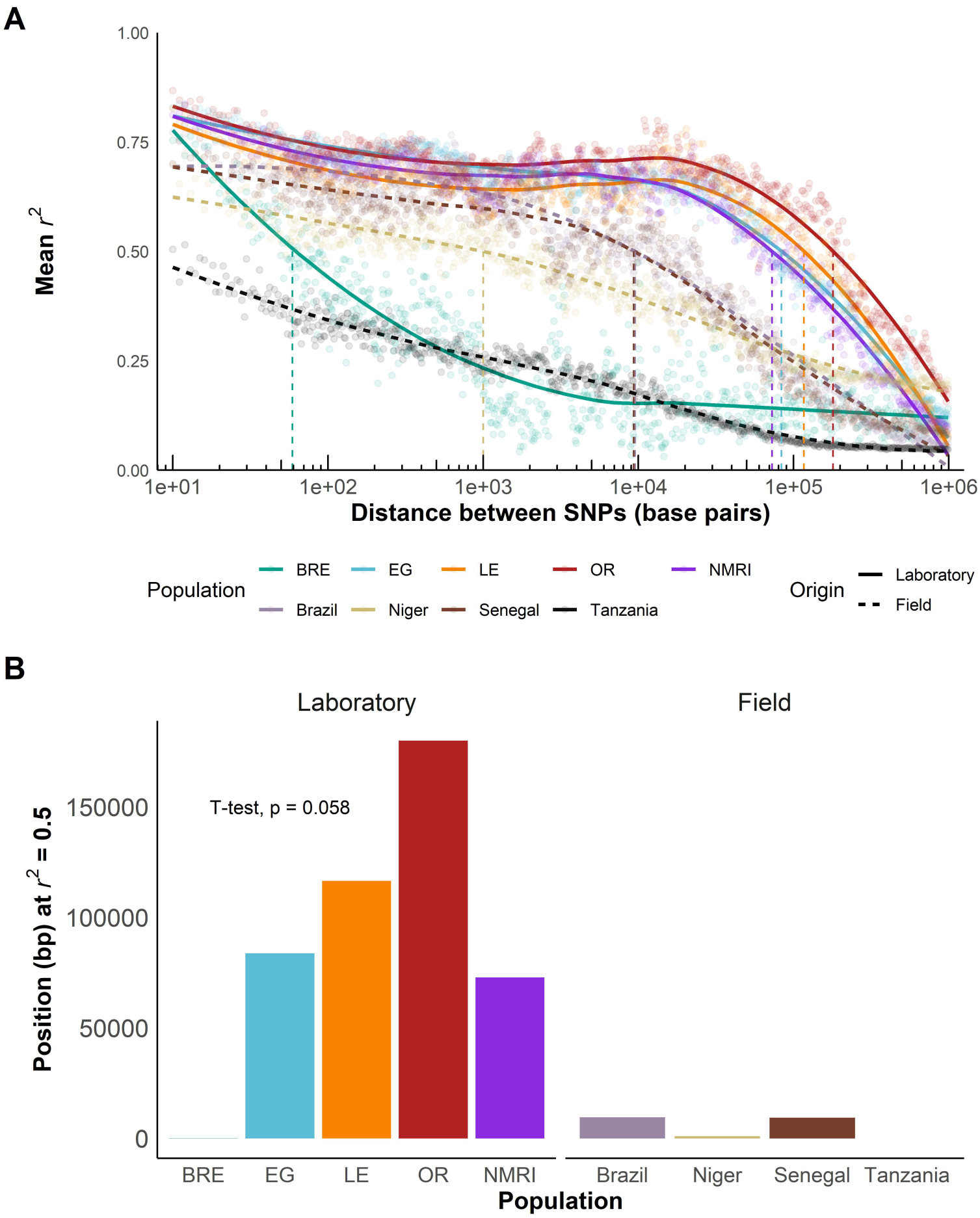


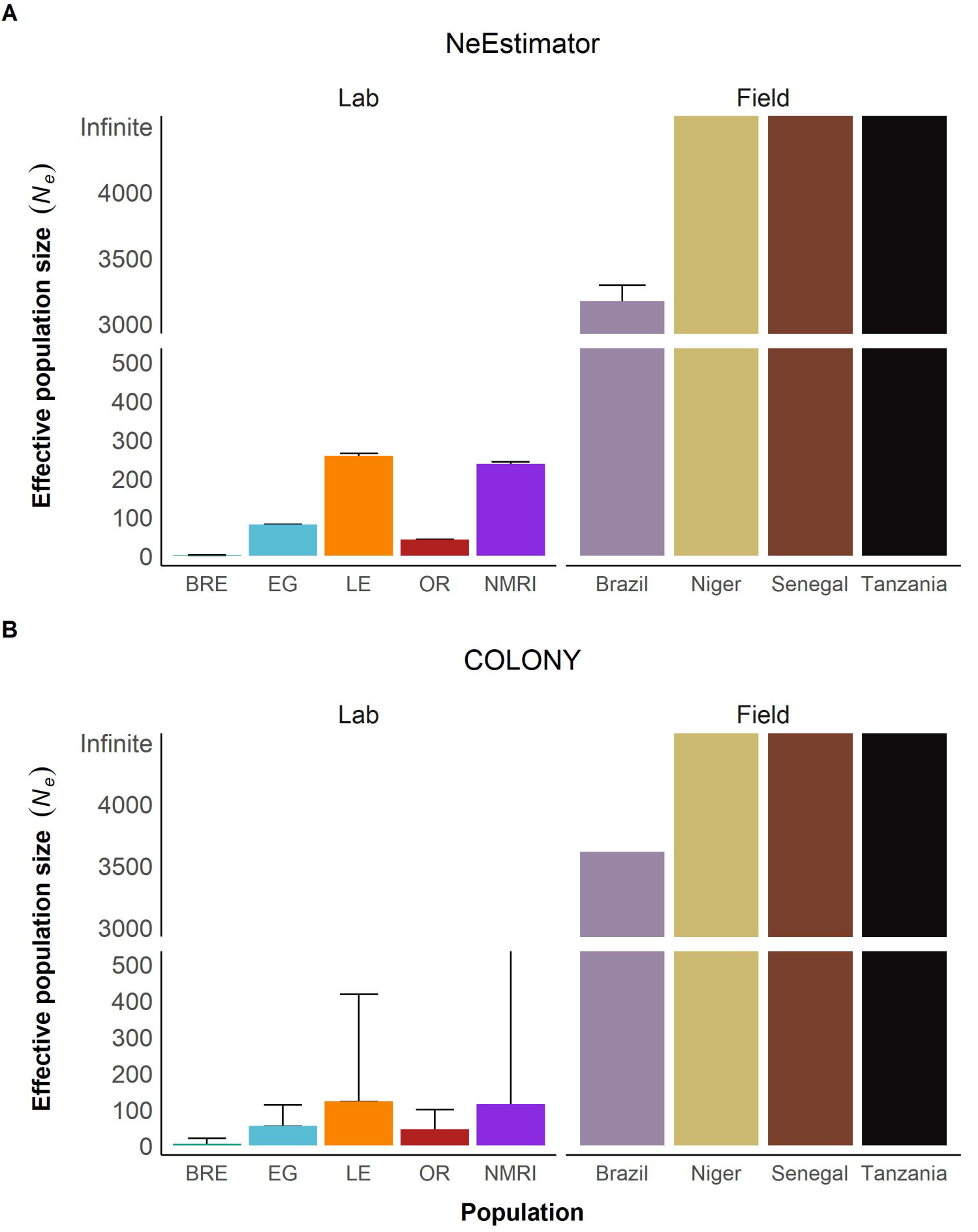


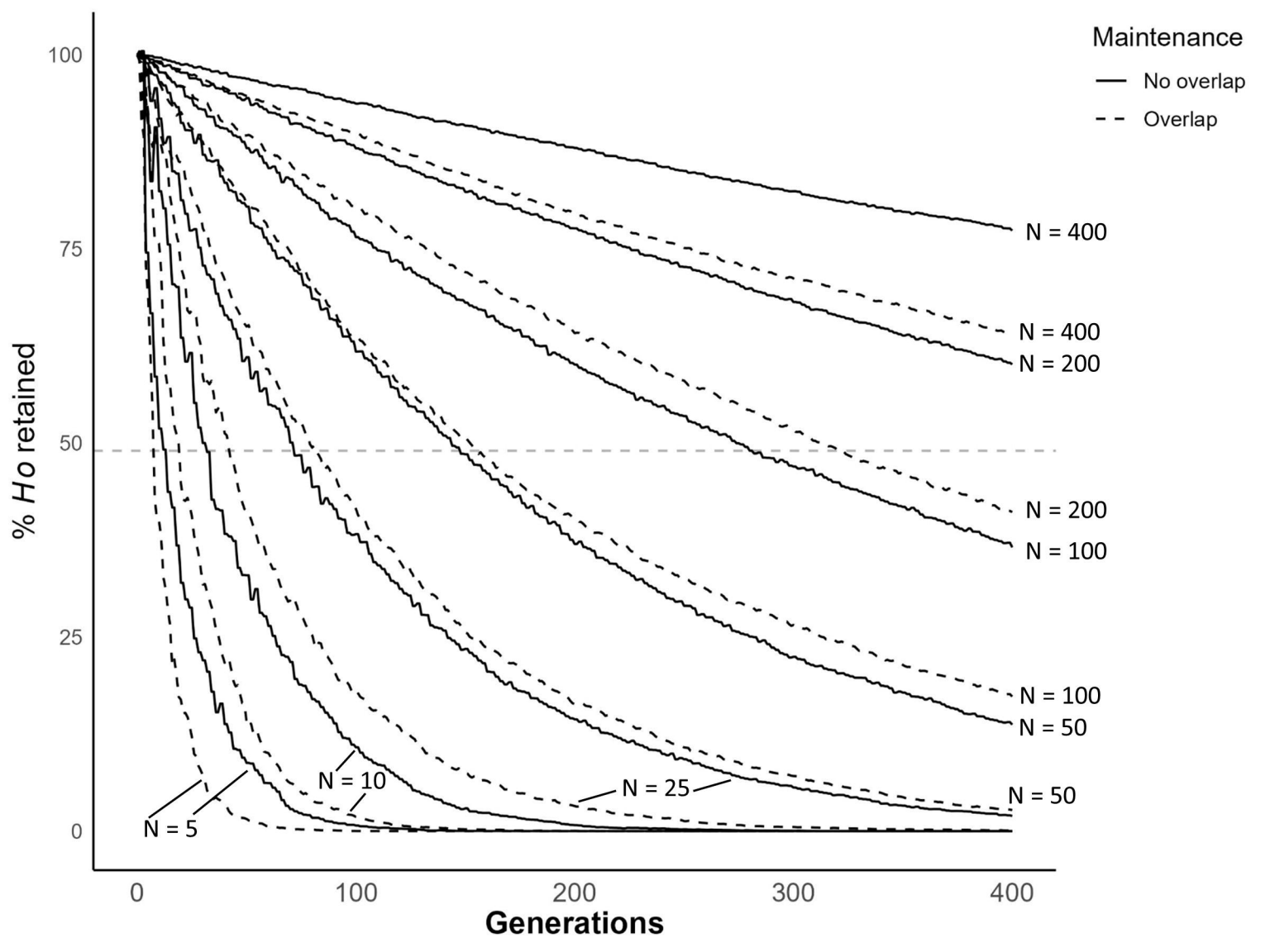



A
Average nucleotide diversity (π)
0.04
0.03
0.02
0.01
0.00
BRE
EG
LE
OR
NMRI
1
2
3
4
5
6
7
Chromosome
B
Laboratory
Field
Kolmogorov-Smirnov, p < 0.001
Average nucleotide diversity (π)
0.003
0.002
0.001
0.000
BRE
EG
LE
OR
NMRI
Brazil
Niger
Senegal
Tanzania
Population


A
Laboratory
Field
Wilcoxon test, p = 0.111
Tajima's D
1.0
0.5
0.0
-0.5
-1.0
BRE
EG
LE
OR
NMRI
Brazil
Niger
Senegal
Tanzania
Population
B
Kolmogorov-Smirnov test, p < 0.001
ECDF
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Allele frequency
Origin
Lab
Field
Population
BRE
EG
LE
OR
NM
Brazil
Niger
Senegal
Tanzania


A
1.00
0.75
0.50
0.25
0.00
Mean r^2
1e+01
1e+02
1e+03
1e+04
1e+05
1e+06
Distance between SNPs (base pairs)
Population
BRE
EG
LE
OR
NMRI
Brazil
Niger
Senegal
Tanzania
Origin
Laboratory
Field
B
Laboratory
Field
T-test, p = 0.058
Position (bp) at r^2 = 0.5
150000
100000
50000
0
BRE
EG
LE
OR
NMRI
Brazil
Niger
Senegal
Tanzania
Population


A
NeEstimator
Lab
Field
Infinite
4000
3500
3000
500
400
300
200
100
0
Effective population size (N_e)
BRE
EG
LE
OR
NMRI
Brazil
Niger
Senegal
Tanzania
B
COLONY
Lab
Field
Infinite
4000
3500
3000
500
400
300
200
100
0
Effective population size (N_e)
BRE
EG
LE
OR
NMRI
Brazil
Niger
Senegal
Tanzania
Population


Maintenance
No overlap
Overlap
100
75
50
25
0
% Ho retained
0
100
200
300
400
Generations
N = 400
N = 400
N = 200
N = 200
N = 100
N = 100
N = 50
N = 50
N = 25
N = 10
N = 5