Pooled, Long-read Sequencing for Structural Variant Characterization in Schistosome Populations

Shalini Nair [1], Xue Li [1], Timothy J.C. Anderson [1], Roy N. Platt [1,*]

[1]Disease Intervention and Prevention Program, Texas Biomedical Research Institute, San Antonio, TX, USA

*Corresponding author: E-mail: rplatt@txbiomed.org.

Accepted: June 18, 2025

Abstract

Pooled sequencing provides a rapid cost-effective approach to assess genetic variation segregating within populations of organisms. However, such studies are typically limited to single nucleotide variants and small indels (≤50 bp), and have not been used for structural variants (SVs > 50 bp) which impact large portions of most genomes and may significantly impact phenotype. Here, we examined SVs circulating in five laboratory populations of the human parasite *Schistosoma mansoni* by generating long-read sequences from pools of worms (92–152 per population). We were able to identify and genotype 17,446 SVs, representing 6.5% of the genome despite challenges in identifying low-frequency variants. SVs included deletions (n = 8,525), duplications (n = 131), insertions (n = 8,410), inversions (n = 311), and translocations (n = 69) and were enriched in repeat regions. More than half (59%) of the SVs were shared between ≥4 populations, but 12% were found in only one of the five populations. Within this subset, we identified 168 population-specific SVs that were at-or-near fixation (>95% alternate allele frequency) in one population but missing (<5%) in the other four populations. Five of these variants impact the coding sequence of six genes. We also identified eight SVs with extreme allele frequency differences between populations within quantitative trait loci for biomedically important pathogen phenotypes (drug resistance, larval stage production) identified in prior genetic mapping studies. These results demonstrate that long-read sequence data from pooled individuals is a viable method to quickly catalogue SVs circulating within populations. Furthermore, some of these variants may be responsible for, or linked to, regions experiencing, population-specific directional selection.

Key words: *Schistosoma mansoni*, nanopore, population genetics, parasite variation, quantitative trait loci.

Significance

Structural variants (SVs) are large genomic variants that are frequently overlooked despite being the largest source of genetic variation within a population. This is because large SVs are expensive and difficult to genotype relative to single nucleotide variants or small indels, so are typically overlooked in population studies. This study attempts to solve these problems by using pooled samples and long-read sequencing to survey SVs circulating in five laboratory populations of the human parasite, *Schistosoma mansoni*. We were able to identify 17,446 SVs that impact 6.5% of the genome. A number of these SVs may be linked to population-specific adaptations. We also found eight SVs that were associated with known parasitic traits from previous studies. This work highlights the value of long-read sequencing of pooled samples to document genetic diversity and provides a new method for exploring the role of SVs in parasite evolution and pathogenicity.

Introduction

Structural variants (SVs), including insertions, deletions, duplications, translocations, inversions, and fusions >50 bp (Alkan et al. 2011), represent a significant source of genetic variation within populations (Sudmant et al. 2015). For example, in humans, two individuals may differ at 0.1% of their genomes when measured by singe nucleotide variants (SNVs), but by up to 1.2% due to SVs (Pang et al. 2010). Furthermore, SVs are frequently associated with major phenotypic changes. The textbook example of the peppered moth (*Biston betularia*) which underwent a color change from white to black in response to coal pollution and bird predation during the industrial revolution was driven by a ~22-kb SV that altered expression of the *cortex* gene (Van't Hof et al. 2016). Other SVs have been linked to disease phenotypes (Weischenfeldt et al. 2013) for multiple human traits including cancers (Li et al. 2020) and autism (Brandler et al. 2018), while in pathogen populations SVs often play roles in drug resistance and immune evasion (Nair et al. 2007; Nair et al. 2008; Cheeseman et al. 2016).

Despite their clinical and evolutionary significance, large SVs are often overlooked in population genetic studies of non-model organisms. This is primarily due to challenges in sequencing and genotyping. SVs are typically larger than the reads generated by short-read sequencing technologies, requiring indirect inference from secondary features of mapped short reads (Gong et al. 2021). While these methods have achieved some success, they can suffer from relatively low accuracy and precision (Sibbesen et al. 2018). Consequently, nearly 70% of all SVs may be missed (Ebert et al. 2021), representing a significant blind spot in the study of within-species variation.

Rapid improvements in long-read sequencing technologies by Oxford Nanopore Technologies (ONT) and Pacific Biosciences now allow generation of reads above 100 kb in length (Logsdon et al. 2020). However, these methods are still expensive relative to short-read technologies and require large amounts of intact DNA from fresh tissue for best results. The central aim of this study was to evaluate the efficacy of pooled long-read sequencing for rapid, cost-effective characterization of SVs segregating within populations. Use of pools has several advantages. First, single individuals may not provide sufficient DNA for nanopore sequencing from fresh tissue, but pools provide ample material. We note that Kim et al. (2024) were able to generate long-read assemblies from single flies, using 35-ng of whole DNA, while high-quality data may be recovered from DNA generated by whole genome amplification (WGA; Lee et al. 2023; Agyabenga-Dadzie et al. 2025). However, nanopore sequencing from multiple individuals would be needed to assess population allele frequencies adding to costs, and WGA has potential to introduce biases and artifacts due to variable coverage, amplification errors,

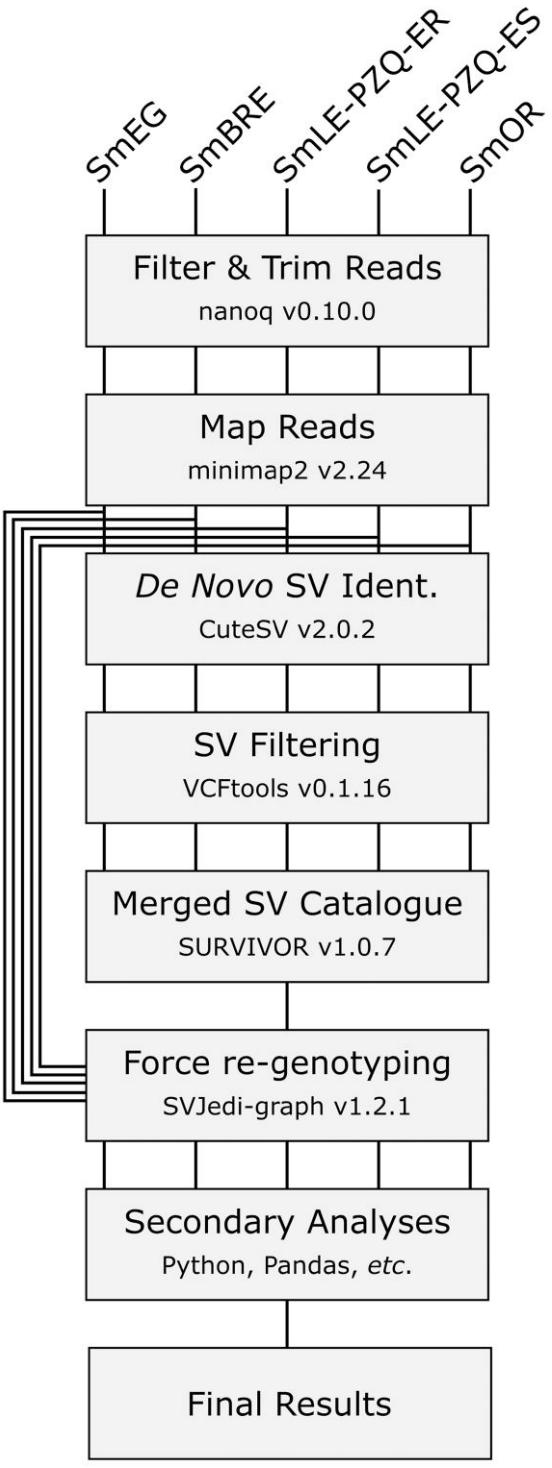

Fig. 1. Bioinformatic methods overview. Long-read sequence data were generated for five *Schistosoma mansoni* populations in order to describe segregating structural variants (SVs). Initially, we identified SVs in each population after filtering and mapping reads to the SM_V10 genome assembly. These initial SV calls were filtered and then merged into a single catalogue representative of all SVs within the *S. mansoni* populations we queried. We used that catalog of SVs to re-genotype the populations at each SV site in the catalogue.

and concatemers (Sabina and Leamon 2015), or to amplify contaminants along with the target DNA (Thoendel et al. 2017). Second, nanopore sequencing of pools of individuals has the added benefit of documenting SVs circulating within the population, similar to the use of pooled short-read sequencing to evaluate SNV population frequencies (Schlötterer et al. 2014; Czech et al. 2024). To date, this approach has not been employed for long-read sequencing and SV characterization.

SVs can have significant impacts on parasite phenotypes. For example, most SVs in the malaria parasite (*Plasmodium falciparum*) populations are at low frequency and likely neutral or deleterious (Cheeseman et al. 2016). However, high frequency copy number variants and duplications in drug resistance genes (Ravenhall et al. 2019) clearly demonstrate the possibility for positive selection to act on SVs. In the protozoan parasite *Leishmania* spp., chromosomal copy number is highly variable and involved in adaptation (Dumetz et al. 2017). Along these lines, SVs are key drivers of diversity in *Giardia*. Comparisons between lines show that one-third of the genome falls within regions affected by SVs. Of particular interest is the expansion and contraction of the variant-specific protein gene family responsible for host immune evasion (Pollo et al. 2020).

We focus on trematodes in the genus *Schistosoma* which includes blood-feeding flukes that parasitize humans and other mammalian hosts. These parasites are diploid, with a ZW sex determination and ~400-Mb genomes (Berriman et al. 2009; Protasio et al. 2012; Buddenborg et al. 2021). *Schistosoma* larvae (miracidia) infect snail intermediate hosts, where they clonally reproduce to generate motile larvae (cercariae, second-stage larvae). These penetrate the skin of the mammalian host, migrate through the circulatory system, and settle in the mesenteric veins. There, mated female worms release hundreds of eggs per day (Moore and Sandground 1956). Eggs are expelled in feces or urine, restarting the life cycle (Nelwan 2019). However, many eggs also become lodged in various organs, resulting in pathology. Schistosomiasis affects >100 million people in the Global South, with the majority of the burden in sub-Saharan Africa (Steinmann et al. 2006).

S. mansoni is particularly amenable to laboratory maintenance (Hackett 1993), and several laboratory populations exist that vary in their geographic origin or vary in biomedically important phenotypes. The genetics of wild and laboratory schistosomes have been extensively studied using SNVs (Berriman et al. 2009; Protasio et al. 2012; Clément et al. 2013; Le Clec'h et al. 2018; Chevalier et al. 2019; Buddenborg et al. 2021; Le Clec'h et al. 2021a; Platt et al. 2022; Vianney et al. 2022) and previous work has shown that *S. mansoni* laboratory populations are diverse when measured at the single nucleotide level (Jutzeler et al. 2024). However, little is known about genetic diversity at the level of SVs in laboratory or field populations.

Our study had three goals; (i) to catalogue SVs circulating in *S. mansoni* laboratory populations using long-read sequencing of pooled individuals as a proof-of-concept; (ii) to identify population-specific SVs that may reflect adaptation (iii) to examine genome regions known to underlie parasite phenotypes for the presence of SVs. We have previously used genome-wide association (GWAS) and linkage studies to map quantitative trait loci (QTL) associated with schistosome phenotypes such as drug resistance or larval stage production (Anderson et al. 2018; Le Clec'h et al. 2021a). These mapping studies typically use short-read sequence data that may miss large SVs. SVs present in QTL regions could be an overlooked source of variation contributing to parasite phenotypes, as is the case in other species (Chakraborty et al. 2018). To accomplish these goals, we sequenced pools of parasites from five *S. mansoni* laboratory populations to describe 17,446 large SVs circulating in these populations. Our results show that SVs can impact up to 6.5% of the genome, including 56 genes, with some populations harboring population-specific variants ($n = 168$) that are at-or-near fixation (<95% AF) which may have evolutionary significance. Moreover, we found that large SVs overlap with known QTLs associated with parasite phenotypes. These observations underscore the importance of SVs in shaping genetic diversity within schistosome populations and demonstrate how pooled nanopore sequencing can provide an efficient rapid approach for assessment of segregating SVs.

Results

Sequencing Results

We generated long-read data with ONT sequencing platform for five laboratory populations of *S. mansoni*; SmBRE, SmEG, SmOR and two populations recently selected from a single population SmLE-PZQ-ER and SmLE-PZQ-ES (Table 1). Instead of sequencing a single individual from each population, we generated libraries from 92 to 152 pooled male and/or female adult worms per population. Pooling individuals allowed us to isolate larger quantities of high-molecular-weight DNA, which is essential for long-read sequencing. This approach also offers the advantage of identifying the presence and frequency of variants across multiple individuals, providing a more accurate reflection of the SVs present within each laboratory population compared to sequencing one or a few individuals at a time. The total amount of whole genomic DNA extracted from these pools ranged from 2 to 5 µg.

Across the five laboratory populations, we generated a combined total of 11.9 million reads, resulting in 66.7 gigabases of filtered data. The number of reads, genome

Table 1 Populations examined

Population	Origin	Approx Est. Date	Num. ♀/♂	Features	Citation
SmBRE	Brazil	1975	152/0	low cercarial shedding	Le Clec'h et al. (2021a)
SmEG	Egypt	1980 to 1990	150/0	African origin	Hassan et al. (2003)
SmLE-PZQ-ER	Brazil	2022	140/0	Resistant to Praziquantel, high cercarial shedding, larger sporocysts, higher infectivity to mice	Le Clec'h et al. (2021a)
SmLE-PZQ-ES	Brazil	2022	65/65	Sensitive to Praziquantel, high cercarial shedding, larger sporocysts, higher infectivity to mice	Le Clec'h et al. (2021a)
SmOR	Puerto Rico	1971	92/0	Resistant to Oxamniquine	Valentim et al. (2013)

Table 2 Summary of long-read seqeunce data, genome coverage, and de novo SV identification

Population	Extraction Method	Num. Filt. Reads (1e6)	Num. Filt. Bases (Gb)	N50 Filt. Reads (bp)	Max Filt. Read Len. (kb)	Mapping Rate	Mean Genome Cov	Num. Raw SV Calls	Num. Filt. SV Calls
SmBRE	NanoBind	2.16	13.73	16,145	193.07	97.96%	33.03	34,805	25,117
SmEG	Monarch	1.40	9.32	13,681	236.94	98.25%	22.14	23,475	18,408
SmLE-PZQ-ER	NanoBind	1.33	7.00	12,071	447.35	97.52%	16.93	19,884	17,694
SmLE-PZQ-ES	Monarch	3.56	22.57	13,259	198.21	93.52%	50.96	44,031	23,776
SmOR	Monarch	3.41	14.12	8,668	168.27	90.33%	31.11	31,624	23,980

Table 3 Summary statistics for each of the major SV types after filtering and merging from all five laboratory populations

	Count	Mean (bp)	Median (bp)	Min (bp)	Max (bp)	Total bp	Count
Deletions (DEL)	8,525	−1,475	−441	−161,688	−63	12,576,321	8,525
Duplications (DUP)	131	9,798	4,504	73	98,879	1,283,568	131
Insertions (INS)	8,410	1,227	427	−179	21,343	10,320,844	8,410
Inversions (INV)	311	5,745	864	−4,424	199,209	1,786,834	311
Translocations (TRA)	69	Na	Na	Na	Na	Na	69

coverage, and N50 values varied between populations, as shown in Table 2. N50 values ranged from 8,668 bp (SmOR) to 16,145 bp (SmBRE), with maximum mapped read lengths reaching up to 447 kb (SmLE-PZQ-ES). The data quantity and quality were comparable between the Monarch and Nanobind DNA extraction methods, however, we prefer the Monarch kits, because they are readily available. Mean genome coverage varied from 16.93× (SmLE-PZQ-ER) to 50.96× (SmLE-PZQ-ES). All sequence data is accessioned under NCBI BioProject PRJNA1206205.

Identifying and Genotyping SVs

We identified 153,819 raw SVs and small indels (<50 bp) across all populations during our initial round of genotyping (Fig. 1), with counts ranging from 19,884 in SmLE-PZQ-ER to 44,036 in SmLE-PZQ-ES (Table 2). After applying a read depth filter (DP > 10), 4,821 SVs were removed, with the majority (43.5%, $n = 2,095$) being removed from SmLE-PZQ-ER, which had the lowest median genome coverage.

The combined impact of the DP > 10 and QUAL > 10 filters reduced the dataset from 153,819 to 108,975 total SVs (29.2% reduction). We applied these filters for two reasons. First, removing variants with low QUAL reduces false positive SV calls. Second, the next step in our analysis involved combining filtered variants across all populations. This means that a low-frequency variant removed due to a low QUAL score in one population might be at high frequency, have a high QUAL score, and be retained in another population. In such cases, the variant would still be included in downstream analyses.

The 108,975 variants from the five populations were merged into a single catalogue in a step that combined overlapping variants and removed any remaining small indels (<50 bp, $n = 70,216$) leaving 17,565 large, unique SVs. A basic summary of the counts and sizes of each of the major SV types is shown in Table 3. We started with 17,565 variants, but dropped an additional 119 overlapping SVs that occurred at the same position in the genome, leaving 17,446 SVs. Of these 97.1% (16,935) were insertions and deletions, while 311 (1.8% were inversions and

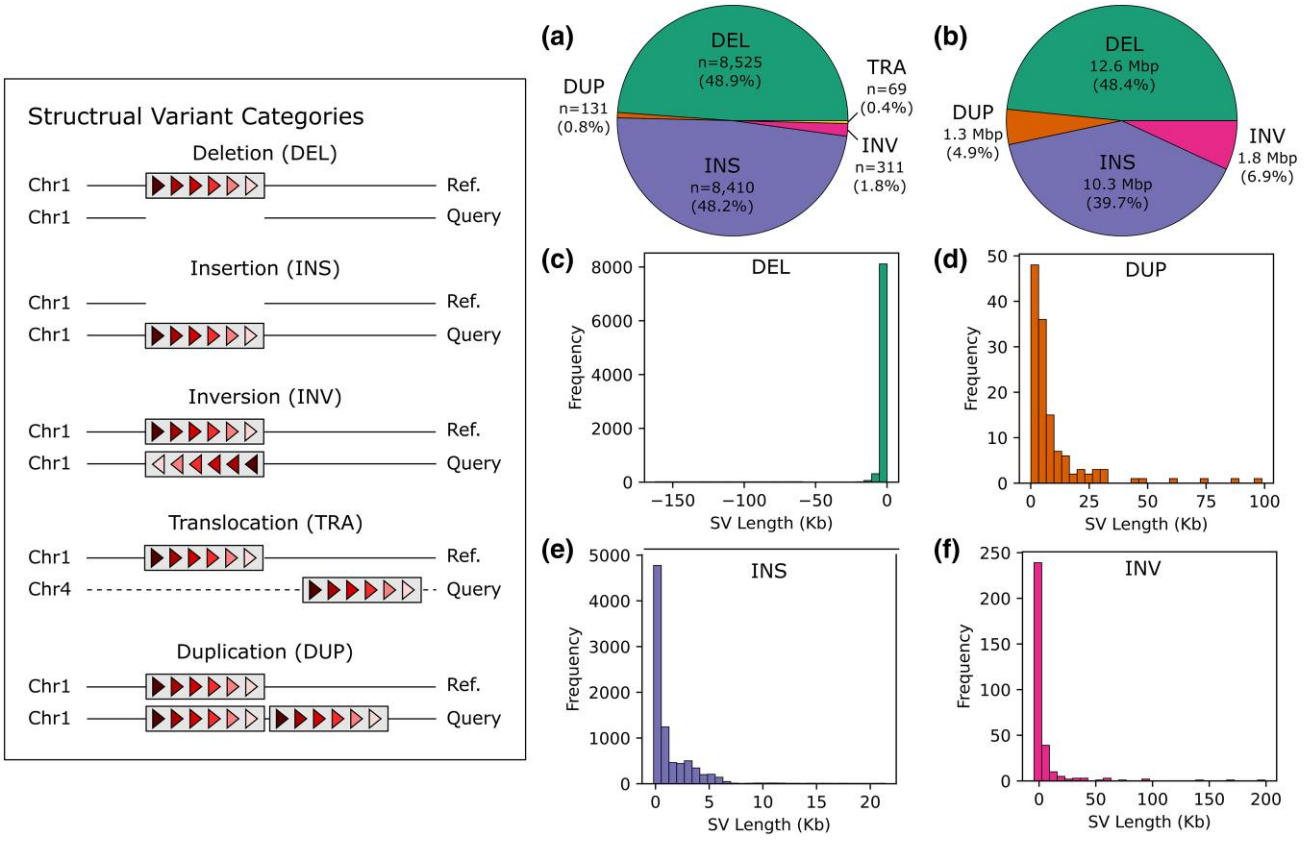

Fig. 2. Abundance and size distributions of SV**s in** five *Schistosoma mansoni* populations. We examined five major types of SVs in total including; deletions ("DEL"), insertions ("INS"), inversions ("INV"), translocations ("TRA") and duplications ("DUP"). a) Most SVs were insertions or deletions, and only ~3% were duplications, transversions, or translocations. b) Similarly, the genomic impact in terms of bases affected by SVs is biased toward insertions and deletions. Translocations often occurred between chromosomes, or between locations on the same chromosome so they were excluded from this subfigure. As a group, the size range for c) deletions, d) duplications, e) insertions, and f) inversions varied between 50 and 199,209 bp.

69 (0.4%) were translocations (Fig. 2). Combined, all SVs identified in these analyses impact 25.97 Mb representing 6.53% of the 397.9-Mb *S. mansoni* genome.

We re-genotyped each population using the catalogue of 17,446 SVs that span the entire genome including autosomes, sex chromosomes, and the mitochondria. Allele frequencies were calculated for each SV at the population level (Fig. 3a). The alternate allele frequency distributions were similar between populations and variant types. Completely fixed variants, where the alternate allele is at 0 or 100% frequency, were particularly abundant, representing 23.5% to 50.2% of any SV type and population combination. Segregating variants were evenly distributed at lower frequencies across each SV type and population combination.

SV Distributions Among Populations

We examined the distribution of SVs among each of the five laboratory populations. Determining absence of an SV is problematic because each population differed in read depth and genome coverage. This affects precision and our ability to confidently state that a variant is missing from a population. To circumvent this issue, we treated all variants at <5% in a population to be "low frequency" rather than absent. Using this criterion, we were able to isolate 1,648 SVs (11.3%) that were at >5% in a single population (Fig. 3b). By comparison 5,349 (36.8%) of all autosomal SVs were at >5% frequency in all populations. While most of the 1,648 population-specific SVs were at low-to-moderate frequencies, 168 of these population-specific SVs are at-or-near fixation (>95% alternate allele frequency). All the population-specific SVs were found in three populations; 114 in SmOR, 46 in SmBRE, and 8 in SmEG (Fig. 4, Table S1). We did not identify any SVs that were specific to SmLE-PZQ-ER and SmLE-PZQ-ES, consistent with the recent divergence of these two populations (Le Clec'h et al. 2021a). However, when these to populations were combined into a single "SmLE" population we did identify 17 population-specific variants (Fig. S1).

We examined the relationships between parasite populations using SV allele frequencies using a principal component analysis (PCA). To focus on autosomal variants and

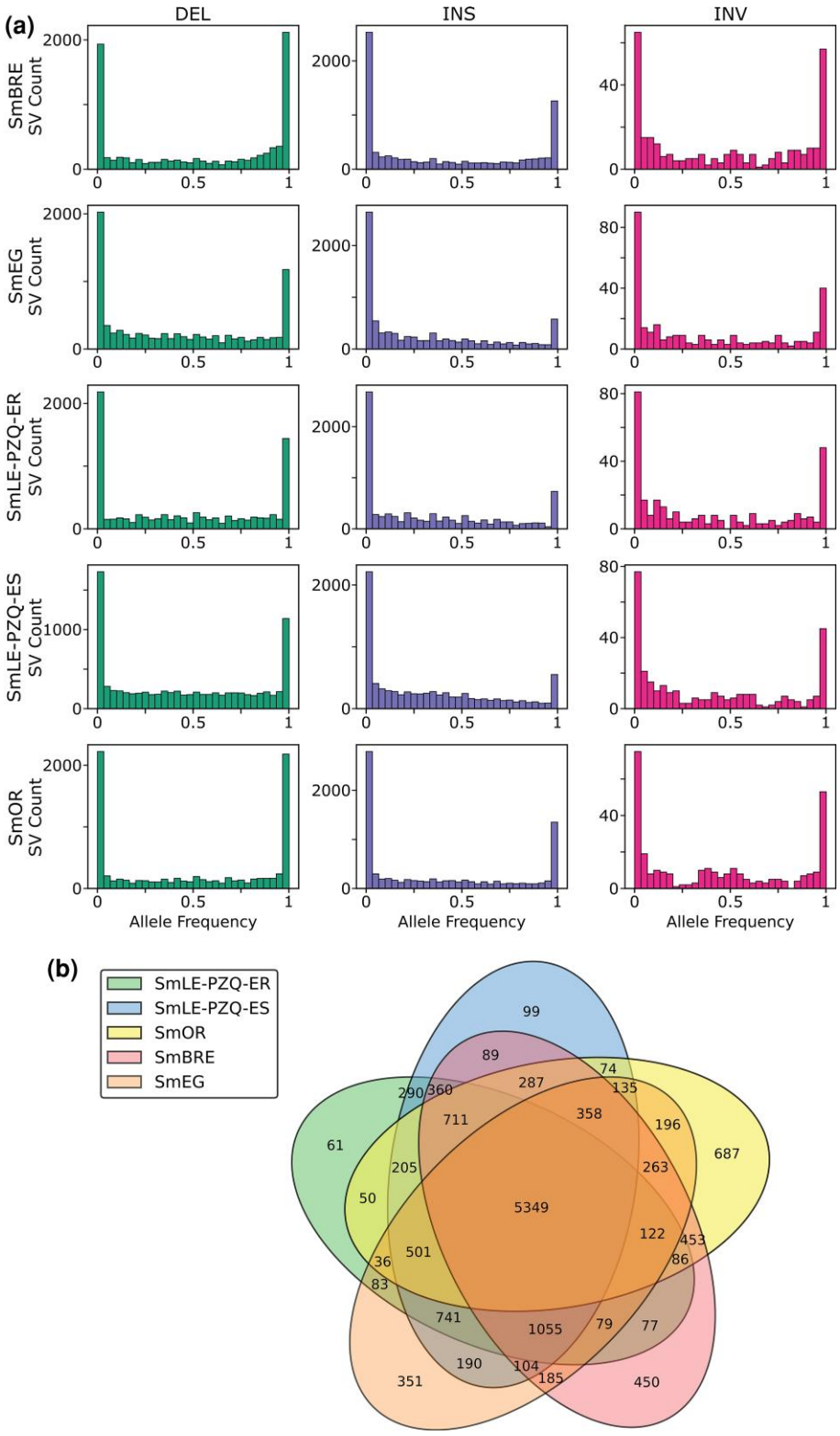

Fig. 3. SV distributions in *Schistosoma mansoni* populations. a) Allele frequency (AF) distributions for each laboratory population of *S. mansoni* broken down by SV type. The AF distributions were relatively consistent for each population and SV type. Most SVs were at-or-near fixation (AF < 0.05 | AF >0.95). All SVs and allele frequencies are called relative to the reference. For example, a deletion specific to the reference genome is identified as a fixed insertion in the other populations. This leads to the bias of high frequency SVs shown above. b) A Venn diagram showing the distribution of SVs among five laboratory populations of *S. mansoni*. SVs were scored as "present" if they were present in the population with an alternate allele frequency >5%.

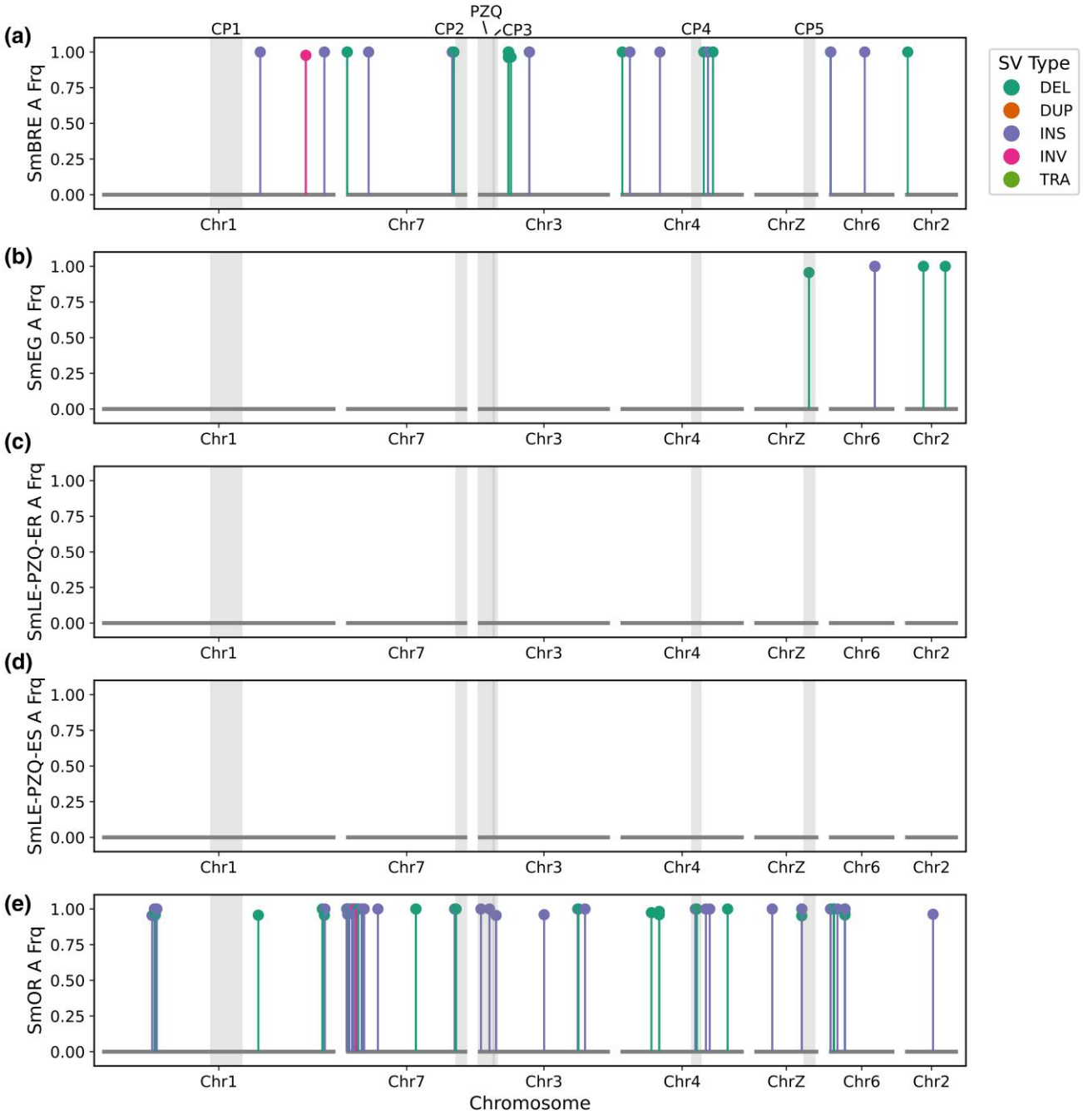

Fig. 4. Population-specific, SVs distribution across the genome. Population-specific variants are at >95% in a single population and <5% in the other four other populations. The distribution of population-specific SVs is not uniformly distributed across the genome. Grey boxes indicate parasite QTLs identified in previous studies and includes five QTLs associated with cercarial production "CP" and a praziquantel resistance QTL (PZQ). We did not identify any population-specific SVs in SmLE-PZQ-ER and SmLE-PZQ-ES. These populations were only recently established from an SmLE source population. Fig. S1 shows the results of this analysis after combining the SmLE populations.

avoid potential biases from skewed sex distributions, we removed 3810 variants located on the Z chromosome, W specific-region, and mitochondria, resulting in a final dataset of 13,656 SVs. Nearly two-thirds of the variation in the PCA (Fig. 5) was explained by the first two principal components (PCs): principal component 1 (36.79%) and principal component 2 (27.88%). The populations were evenly distributed in PC space, with the exception of SmLE-PZQ-ER and SmLE-PZQ-ES, which nearly overlapped. This overlap is not unexpected, as these two populations were only recently derived from a common SmLE-PZQ-R progenitor stock (Le Clec'h et al. 2021a).

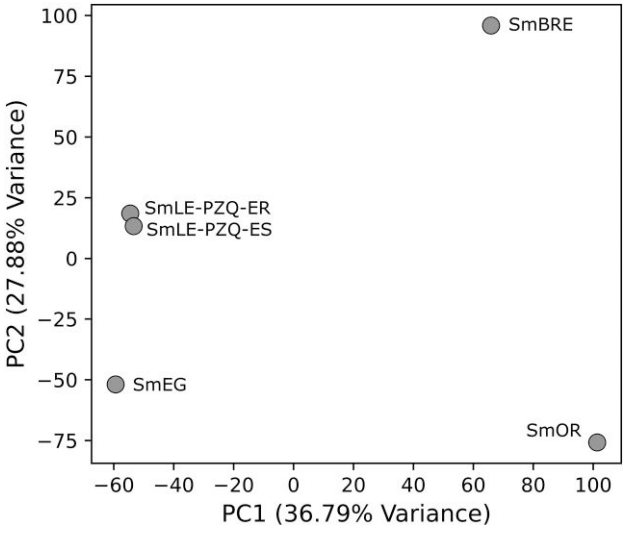

Fig. 5. Population structure of five *Schistosoma mansoni* laboratory populations using allele frequencies from 17 446 SVs. SVs were scored as present or absent in a population if the allele frequency was >5%. Components 1 and 2 account for 36.79 and 27.88% variation in the dataset. SmLE-PZQ-ER and SmLE-PZQ-ES populations cluster together. These populations were derived from a single SmLE parent population by selecting for parasites sensitive or resistant to the praziquantel anthelminthic drug.

SV Impact on Gene Regions

We examined SVs in coding sequences and regulatory regions which have the potential to affect parasite phenotypes (Fig. 6). We identified 168 population-specific, at-or-near fixed (>95% AF) SVs, of which 70 (41.7%) overlap with 56 different genes (Table S1). Most of these variants are located in intronic regions ($n = 50$); however, we found population-specific variants affecting six genes within either the UTR or coding sequence. These variants include one inversion, one deletion, and four insertions. The inversion spans approximately 30 kb (SM_V10_2: 3,848,692 to 3,878,697) and contains two genes (Smp_327700 and Smp_327710). Both are identified as secreted proteins by WormBase ParaSite (WBPS19), but little is known about their expression profiles or functions. Similarly, there is limited information available on Smp_158710. We found a deletion which removes part of an intron and 41 bp of exon3 (SM_V10_6: 6,022,561 to 6,022,739) in SmOR. Smp_315070, (a SERPIN domain-containing protein) contains a pair of 577 and 2,293 bp insertions in the 5′ UTR in the SmOR population. Smp_315070, is expressed at low levels in many different tissue types (Wendt et al. 2020). Smp_342090, a low-density lipoprotein receptor class A, contains a 364 bp insertion in the 5′ UTR in SmBRE. This gene is highly expressed in most types of neuronal cells (Wendt et al. 2020). The final gene, Smp_076960, encodes a C2 NT-type domain-containing protein that is expressed in most major cell types (Wendt et al. 2020). There is a 419 bp insertion in the 3′UTR of this gene in SmOR.

SVs Within Known Parasite QTL

Previous work has established a single QTL associated with praziquantel drug resistance (Le Clec'h et al. 2021a; Chevalier et al. 2024) and five QTLs for cercarial production (Le Clec'h et al. 2021b). We converted the coordinates of these loci to the current SM_V10 assembly and examined the SVs within these regions, focusing on variants with strong allele frequency differences (>95%) between the populations analyzed in the original studies (Table 4).

In the single QTL associated with praziquantel drug resistance, we found two variants of interest. While these variants are strongly differentiated between SmLE-PZQ-ER and SmLE-PZQ-ES, they are also found circulating in other *S. mansoni* populations at frequencies ranging from 0 to 100%. Notably, two variants, DEL.SM_V10_3: 2,969,428 to 2,973,340 and DEL.SM_V10_3: 3,194,040 to 3,194,564, impact the intronic regions of Smp_345310 (Transcription factor SOX-13), and Smp_307630 (Zinc finger, TFIIS-type; Pol I subunit A12, C-terminal zinc ribbon), respectively. Among the five QTLs associated with cercarial production, two SVs exhibited $a > 95\%$ frequency difference between SmBRE and the SmLE populations: a 6,146 bp insertion at SM_V10_5: 21925347 between Smp_330830 and Smp_086690 (Protein kinase domain; Serine/threonine-protein kinase, active site; Protein kinase, ATP binding site) and a 76 bp insertion (INS.SM_V10_2: 41902922). In these cases, we used the mean frequency of these SVs in SmLE-PZQ-ER and SmLE-PZQ-ES, which were derived from SmLE.

Repeats and SVs

De novo searches of the genome identified 985 families of repetitive sequences. We could assign only 228 of these to known transposable element (TE) families and 757, were unknown repetitive sequences. More than half (53.3%) of the genome was derived from repetitive sequences, with the largest class of elements falling into the unknown category (27.05%) followed by retrotransposons (24.39%; Fig. 7). Bov-B RTEs and L2s are the most common families of described retrotransposons in the genome. By comparison 84.01% of SV alleles are derived from repetitive sequences and the difference between the two categories (whole genome vs. SV alleles) is driven almost entirely by the expansion of the Gypsy family of long-terminal repeats (LTRs). These are present in 4.1% of the genome but in 23.86% of the SV alleles. The overall repeat content in the SV alleles is significantly greater than in the genome as a whole (two-proportion z-test: $P = 0.0$, $z\text{-score} = 2{,}916$). Allele frequencies of Gypsy-derived SVs are comparable to allele

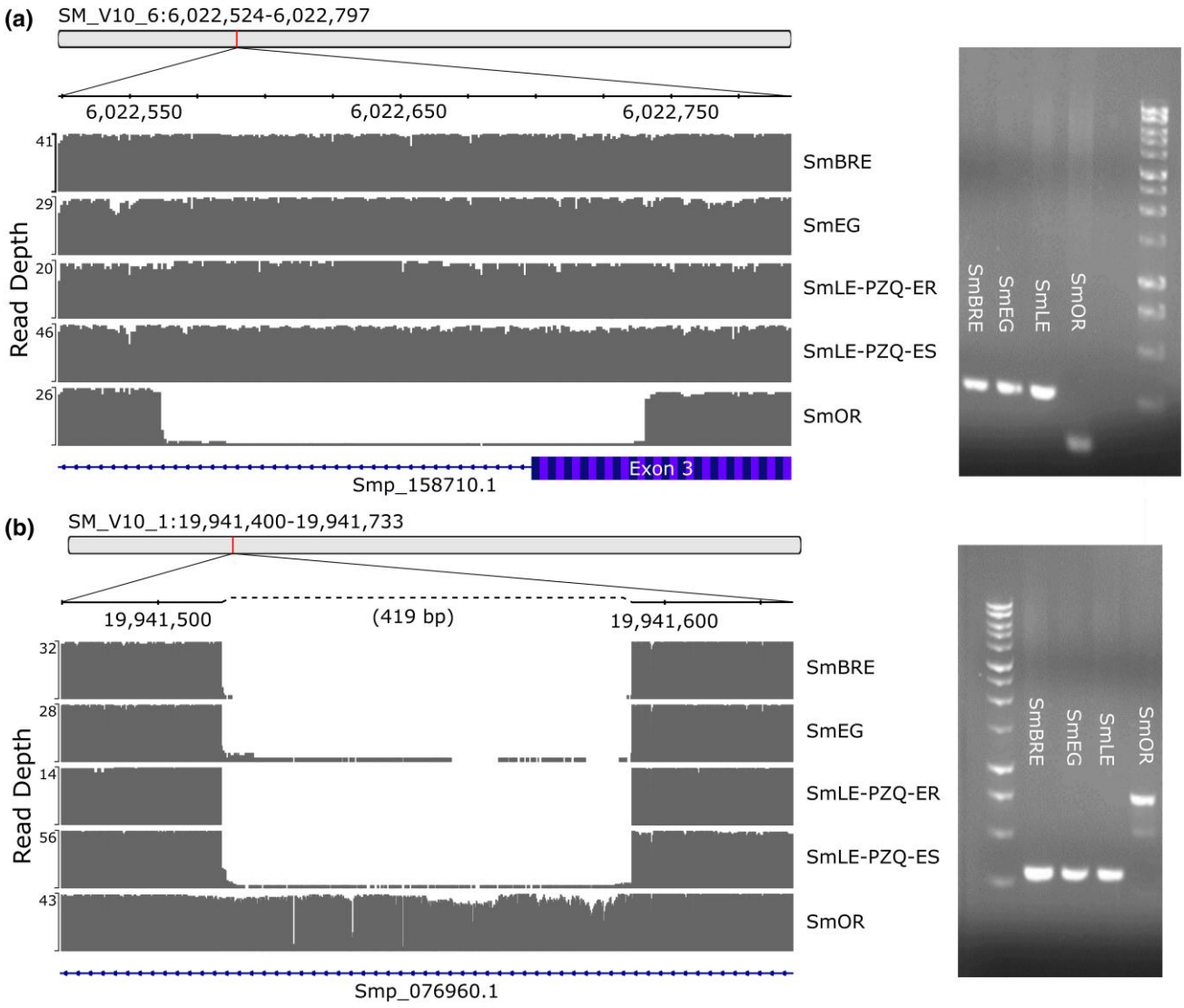

Fig. 6. Examples of SVs within genic regions. a) A 178 bp deletion in a gene (Smp_158710.1) with unknown function is specific to the SmOR population removes a portion of exon 3. b) A 419-bp insertion within the 5′ UTR of a C2 NT-type domain-containing (Smp_076960.1). Both SVs were verified using PCR.

frequencies of all other SVs with mean allele frequencies of 40.8% and 41.2% (two-sided Mann–Whitney *U* test, $P = 0.4994$; Fig. S2a), however the relative age of Gypsy-derived SVs is significantly younger than Gypsy elements in the genome as a whole when measured using Smith–Waterman distances (mean difference = 6.68%, *P*-value < 0.001, 10,000 permutations; Fig. S2b).

SV Validations

We used PCR methods to validate SVs from individuals within each population, focusing on population-specific SVs found within genes or under QTL peaks. These included 11 insertions, 12 deletion, and 2 inversions. We generated primers (Table S2) to successfully amplify 19 of 25 loci. This included 12 population-specific variants, 2 coding variants, and five variants in QTL regions. In general, the PCR results were consistent with the expected (Table 5; Fig. S3). We were able to validate SVs with 90% accuracy, 85% sensitivity, and 93% specificity.

Discussion

The availability of long-read sequencing is rapidly changing our ability to characterize SVs (Mahmoud et al. 2019). In this study, we used long-read sequencing to identify SVs in five laboratory populations of schistosome worms with the goals of cataloguing large SVs in *S. mansoni* laboratory populations and identifying variants that may impact parasite phenotypes. By sequencing parasite pools, rather than individual parasites, we aimed to evaluate frequencies of SVs segregating within each population.

Table 4 QTL associated with parasite phenotypes and their locations in the current *S. mansoni* genome assembly

Phenotype	Original assembly	Original location	SM_V10 location	Populations	SVs	CDS SVs	Variants of Interest
Praziquantel resistance[a]	GCA_000237925.5 (v9)	SM_V9_3: 271315 to 5811857	SM_V10_3:271314 to 5994738	SmLE-PZQ-R, SmLE-PZQ-S	593	12	2
Cercarial production[b]	GCA_000237925.2 (v7)	SM_V7_1: 42010946 to 53482618	SM_V10_1:41070262 to 52541936	SmLE, SmBRE	298	3	2
Cercarial production[b]	GCA_000237925.2 (v7)	SM_V7_2: 43954213 to 47725512	SM_V10_3:5895758 to 7219504	SmLE, SmBRE	74	1	0
Cercarial production[b]	GCA_000237925.2 (v7)	SM_V7_3:3914558 to 5238304	SM_V10_2:41597599 to 45311388	SmLE, SmBRE	246	5	3
Cercarial production[b]	GCA_000237925.2 (v7)	SM_V7_4:27609036 to 30741077	SM_V10_4:26999137 to 30131178	SmLE, SmBRE	64	2	1
Cercarial production[b]	GCA_000237925.2 (v7)	SM_V7_5:19823314 to 23768112	SM_V10_5:18834925 to 22640851	SmLE, SmBRE	237	3	0

[a]Praziquantel resistance loci were described in Chevalier et al. (2024).
[b]Cercarial production loci were described in Le Clec'h et al. (2021b).

Challenges in Identifying Low Frequency and Small Indel Variants in Pooled Samples

Pooled sequencing using short-read, Illumina approaches is now widely used due to its rapid, cost-effective determination of SNP allele frequencies within populations (Lynch et al. 2014). Relative to short reads, long reads are better able to recover indels >10 bp in general (Kosugi and Terao 2024) and specifically insertions >50 bp which are only discoverable 17% to 28% of the time (Delage et al. 2020). So comparable methods for measuring SV frequencies using nanopore long-read sequences from populations could be extremely valuable. However, identifying SVs from pools of individuals presents challenges compared to genotyping single individuals (Layer et al. 2014). We initially identified 19,884 to 44,031 SVs for each of the five laboratory populations. This was reduced to 17,446 SVs after multiple rounds of filtering and merging into a single set of consensus variants. We outline several limitations and opportunities for future experiments.

Most indels are small (<50 bp). We identified 70,216 small indels across all five laboratory populations. These smaller variants outnumber large SVs by nearly 4:1. The bias toward smaller variants is commonly seen in other taxa (Alonge et al. 2020; Trost et al. 2021; Ruggieri et al. 2022). This is not unexpected given that the larger a variant is the more likely it will be to affect functional sequences and be exposed to selection (Chiang et al. 2017). Further, given our pooled sampling strategy it is difficult to differentiate small indels from sequencing artifacts. In general, nanopore sequencing error rates are elevated in homopolymer regions making it difficult to differentiate between true indels and sequencing artifacts (Ye et al. 2025). For example, in humans, error rates in homopolymer regions are 2.6-fold higher than SNVs (Cretu Stancu et al. 2017) and can account for half of all sequencing errors (Delahaye and Nicolas 2021). Bioinformatic tools are available to polish or correct these regions using a variety of techniques from haplotype-based de Bruijn graphs (Liu et al. 2024) to measuring the timing/kinetics of the DNA molecule as it passes through the physical nanopore (Sarkozy et al. 2017). Some appear to be fairly robust (Dohm et al. 2020). What is not clear, though, is how appropriate these methods are for correcting homopolymer reads in pooled samples. Future work could compare in-silico polishing to a known catalog of high-quality variants in homopolymer regions. If successful this technique could help differentiate true SVs from those introduced by sequencing artifacts.

Limitations on Calling Low-Frequency SVs

Most SV tools, and quality metrics like QUAL and GQ, are designed to identify germline mutations that exist as homozygous or heterozygous variants. Pools of individuals defy these expectations since SVs can be present at frequencies ranging from 0 to 100% within a population compared to the 0, 50, and 100% expected in a single diploid individual. QUAL scores assess the confidence that an SV exists at a particular site. In our dataset, even moderate frequency (<27%) variants did not have QUAL scores higher than 10 indicating 90% confidence in a call (Fig. S4). This means that a low-to-moderate frequency variant in population "A" is only present in the final catalogue if it is also present at relatively high frequency in another population(s). As a result, our catalogue under-reports low-to-moderate frequency variants in schistosomes populations.

Sequence Coverage and SV Call Accuracy

Genome coverage also contributes to accuracy in low-frequency SV calls (Yang 2020). The number of SVs was positively correlated with genome coverage (Fig. S5) both before ($R^2 = 0.97$) and after filtering ($R^2 = 0.77$). Further, the ability to detect low-frequency variants with accuracy and precision is directly related to sequencing coverage. Ideally, a

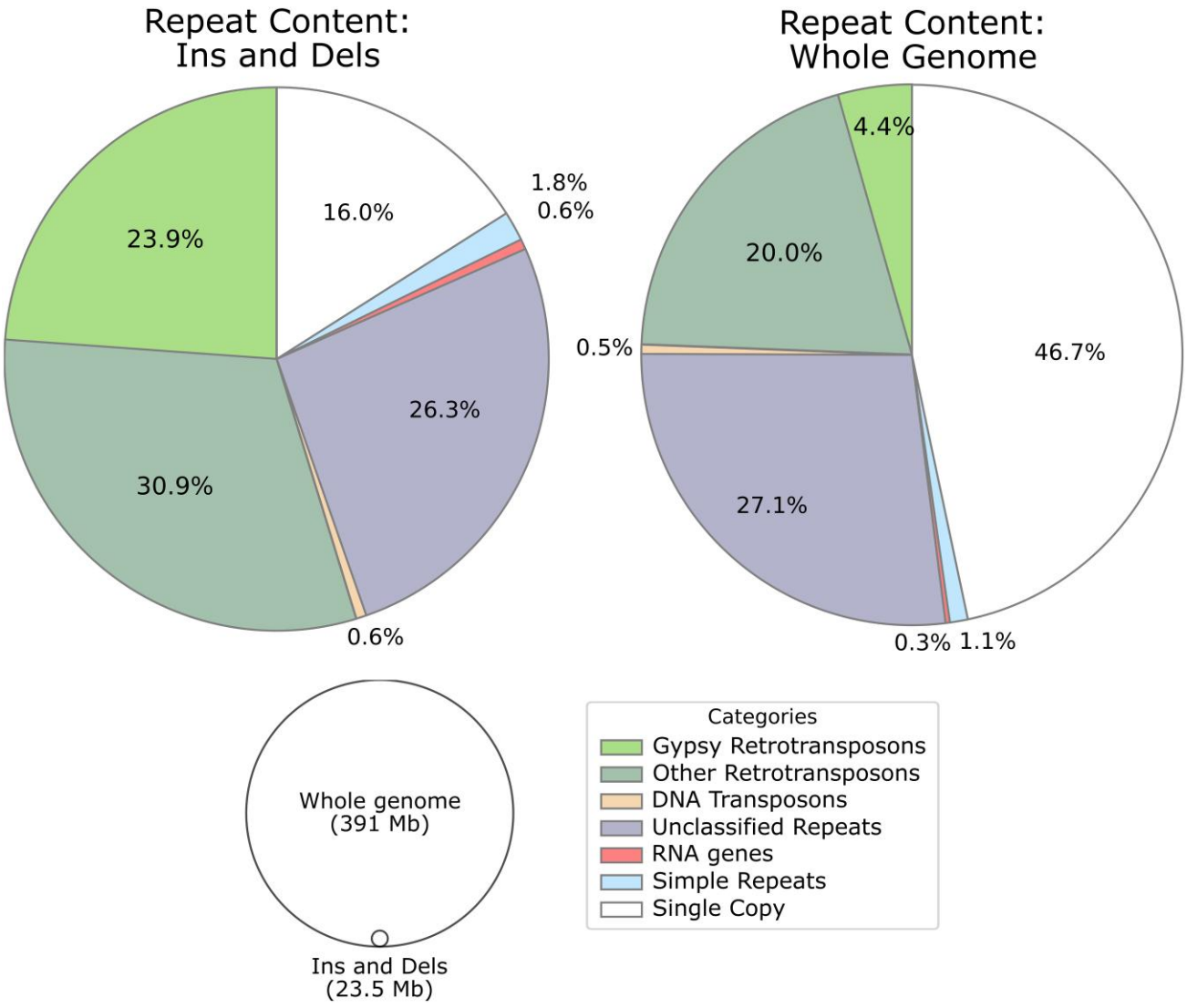

Fig. 7 . SV**s are frequently associated with repetitive sequences.** Around half of the *Schistosoma mansoni* genome is derived from repetitive sequences, but 84% of SVs are either caused by, or found within repetitive regions of the genome. The difference appears to be associated primarily with retrotransposon transposable elements (TEs), primarily those in the Gypsy family.

Table 5 PCR validation of SVs in SmBRE, SmEG, and SmOR.

	n	False neg.	True neg.	True pos.	False pos.	Accuracy (%)	Sensitivity (%)	Specificity (%)
Pop-Specific	12	2	44	20	0	97	91	100
QTL	5	2	6	10	2	80	83	75
Genic	2	0	8	4	0	100	100	100
Overall	19	6	63	34	5	90	85	93

Accuracy = (TP + TN)/(TP + TN + FP + FN), Sensitivity = TP/(TP + FN), Specificity = TN/(TN + FP).

population should be sequenced to a depth greater than the number of individuals (genomes) present in the pool to avoid missing low-frequency variants. For example, in simulated pools of cancer cells, an SV at 20% allele frequency is only detected 31% of the time at 10× coverage (Layer et al. 2014). We estimate that by sequencing 100 pooled worms from the SmLE-PZQ-ES population to 51× coverage, the lowest frequency variants we can score with 95% confidence is 5%. By contrast, we can only be confident in sampling variants present at >15% in SmLE-PZQ-ER due to lower total genome coverage (17×). Our inability to sample SVs at low coverages is compounded by the fact that low-frequency variants are harder to distinguish from false positive and false negative calls using standard quality metrics.

Long-Read Sequencing of Pooled Individuals Can be Used to Identify Large SVs

By pooling parasites from each population, we were able to quickly survey SVs from five *S. mansoni* populations. Even with the limitations discussed above, we were able to identify 17,446 SVs segregating within the laboratory populations. Combined, these SVs, impact ~26 Mb representing 6.5% of the genome. This number is almost certainly an underestimate since low-frequency variants were filtered from the final SV catalogue.

SVs can be used as markers to understand relationships between populations. The majority of SVs where shared between ≥4 populations ($n = 8,096$; 59%) compared to only 1,648 (12%) lineage-specific SVs. We used SV allele frequencies to project the populations in genotypic space using a PCA (Fig. 7). While we are limited to five data points, the patterns are consistent with the history of each laboratory population. SmLE-PZQ-ER and SmLE-PZQ-ES are derivatives of the SmLE population based on their susceptibility to the drug praziquantel (Le Clec'h et al. 2021a). In the PCA, the two populations are effectively overlapping. Further, these two populations combined contain 160 unique variants while each of the other populations contain at least 350 unique, moderate frequency (>5%) SVs (Fig. 3b).

Population-specific SVs Can Affect Coding Sequences

We found 168 population-specific variants within genic regions: these were usually in introns. There are five SVs that directly impact coding regions (CDS) of six genes. Smp_327700 and Smp_327710 are secreted proteins found in a 30-kb inversion. Smp_315070 (a SERPIN domain-containing protein), Smp_158710, and Smp_342090 have unknown function. Smp_076960 (C2 NT-type domain-containing protein) contains a 419 bp insertion in the 5′ UTR. This variant is specific to the SmOR population and appears to be expressed throughout the parasite's life cycle (Harris et al. 2020) and an ortholog is associated with abnormal sleep recovery and lethargy in *C. elegans* (WBPhenotype:0001524; Huang et al. 2017). While it is difficult to directly associate any of these variants directly with a phenotype, these results show that SVs that alter disrupt CDS are present.

Population-specific variants are not randomly distributed throughout the genome (Fig. 4; chi-square test; *P*-value = 1.2e-180). Instead, we found seven windows accounting for 28 Mb of the genome (~7%) that contains 109 of the 168 (64.9%) of all population-specific variants. Within these regions, population-specific variants occur every 3.3 kb (mean) and outside of these regions SVs are distributed every 441.8 kb (mean).

High Frequency SVs are Found in QTL Associated With Parasite Phenotypes

SVs have been associated with phenotypes in many species including horses (Bellone et al. 2023), pigs (Zong et al. 2023), goats (Guo et al. 2021), humans (Hollox et al. 2022), and others (Zhang et al. 2016; Jeffares et al. 2017; Guo et al. 2020; Chen et al. 2023). We have used both linkage mapping and association studies to identify the genetic basis of parasite phenotypes in *S. mansoni* (Anderson et al. 2018; Le Clec'h et al. 2021b; Chevalier et al. 2024). We investigated whether SVs were located within QTL regions for praziquantel resistance or release of larval parasites (cercarial shedding) from the intermediate host snails. We found population-specific SVs of interest in four of six QTLs examined. The presence of these SVs does not necessarily imply they are the causative variants for the observed phenotypes, but these are potential candidates. A transient receptor potential channel gene (TRPM-PZQ) is thought to underlie the praziquantel resistance QTL on chr 3 (Le Clec'h et al. 2021a; Park et al. 2021). However, specific causative SNPs within this gene have not been located. It is possible that SVs outside of the TRPM-PZQ result in altered expression of this gene and modify PZQ-resistance. Interestingly there is a 524 bp deletion in the intronic regions of Smp_345310 (transcription factor SOX-13): this gene shows the strongest association with praziquantel resistance within the QTL region (Le Clec'h et al. 2021a; Chevalier et al. 2024).

Chevalier et al. (2024) used short-read SV calling method to identify two large deletions tightly linked with the praziquantel QTL. These deletions were at SM_V10_3: 2,775,001 to 2,900,000 and SM_V10_3: 3,175,001 to 3,300,000; referred to here as the "upstream" and "downstream" deletions, respectively. We did not recover either of these deletions in the nanopore analysis and only found one deletion that differentiates the two SmLE populations within these QTL regions (SM_V10_3: 3,194,040 to 3,194,564). Mean read depth across the deletions is 0.07× (upstream) and 1.11× (downstream) and both are in the 0.03 and 0.06percentile when compared to the rest of the genome sampled in 125-kb windows. This is consistent with large deletions. By comparison, these same regions are in the 96.2th and 17th percentiles in SmLE-PZQ-ES. Visual inspection of the read alignments is not consistent with an immediate break typically associated with deletions. Instead, genome coverage gradually changes over hundreds of base pairs in regions with high repeat. In the end it is likely that this combined with low coverage in SmLE-PZQ-ER in general and variation in pooled populations likely negatively impact SV calling in these regions.

Repetitive Sequences are Strongly Linked With SVs in *S. mansoni* Populations

We estimate that more than half (53.3%) of the *S. mansoni* genome is derived from repetitive sequences. This is higher than estimates from previous work on *S. mansoni* (45%; Berriman et al. 2009) and *S. haematobium* (43%; Young et al. 2012). The difference between our analysis and

Berriman et al. (2009) could be driven by many factors. For example, Berriman et al. (2009) used RepeatScout (Price et al. 2005) to identify repetitive sequences. Here we used RepeatModer2 (Flynn et al. 2020) which uses a combination of RepeatScout (Price et al. 2005), RECON (Bao and Eddy 2002), LTRharvest (Ellinghaus et al. 2008), and LTR_retreiver (Ou and Jiang 2018).

Most of the repeats (27.1%) from our de novo analysis were unique and did not share significant homology or structural features for classification. Hence, >100-Mb of the 400-Mb *S. mansoni* genome is derived from repeats of unknown origin. Future work should look to provide a complete annotation (Platt et al. 2016; Goubert et al. 2022) and deposit the repeats in a public repository (ex. RepBase; Bao et al. 2015) as a first step in understanding the impact of repetitive sequences on the evolutionary history of *S. mansoni*.

It is apparent that *S. mansoni* SVs are significantly enriched in repetitive sequences [two-proportion *z*-test ($P = 0.0$, *z-score* = 2,916.7)]. While 53.3% of the genome is repetitive, 84% of SVs are caused by, or occur in repetitive sequences. The over-representation of repeat sequences in SVs is consistent with previous work (Domínguez et al. 2020; Reis et al. 2023; Quah et al. 2024; Smeds et al. 2024). Furthermore, we note that highly, repetitive sequences larger than 150 bp are difficult to unambiguously map with short-read sequence data. For example, short reads are unable to map to 10% of the human genome due to large repetitive sequences (Jain et al. 2022). This implies that genotyping SVs with short reads could potentially miss many of the variants associated with repetitive sequences (Kosugi and Terao 2024). Direct comparisons of short vs. long reads for SV calling in humans has shown that 70% to 80% of SVS are called in repetitive regions compared to only 27% to 58% with short reads (Kosugi and Terao 2024) though it should be noted these results are from single individuals rather than population pools. The strong association between repeats and SVs provides a further limitation on our ability to score SVs with short-read Illumina data, which cannot be effectively mapped in repeat regions.

The increase in repeat-related SVs is driven almost entirely by Gypsy endogenous retrovirus which are present in 4.4% of the genome but 23.9% of SV alleles. Structurally, endogenous retroviruses contain a combination of gag, pol, and env genes flanked by two LTRs. The LTRs may range between 100 and 1,000 bp in length and are exact duplicates of each other. Because of their size and sequence similarity, LTRs are hotspots for non-homologous recombination which removes the internal regions of the ERV and one of the LTRs, leaving a behind a single LTR (Hughes and Coffin 2004). The disproportionate number of Gypsy LTRs associated with SVs suggests that they are actively mobilizing in the genome, creating insertions, and being partially removed via non-homologous recombination and generating deletions. Age estimates based on sequence divergence show that Gypsy-derived SVs are significantly younger than Gypsy elements in the genome as a whole (Fig. S2b). This could be caused by a range of phenomena including selection against older insertions or differences in subfamily mobilization rates. In either case, these observations imply that SVs associated with Gypsy mobilization could be a significant driver of lineage-specific evolution in *S. mansoni* laboratory populations.

Conclusions

We used long-read sequencing of pooled parasite populations to identify and describe SVs within five *S. mansoni* laboratory populations. While these methods are limited in their ability to de novo detect low-frequency variants from pooled samples, we were able to identify 17,446 SVs impacting 6.5% of the genome. Furthermore, we identified population-specific variants present at high frequencies and SVs that fall within known parasite QTLS for important phenotypic traits. Such population-specific SVs may reflect the action of selection or determine phenotypic differences between population so are of particular interest. As a proof of concept, this study demonstrates the value of using long-read sequencing and pooled samples to survey SVs within parasite populations. To ensure that the variation generated by SVs is properly accounted for, future work should aim to improve low-frequency variant identification, to evaluate SV frequencies with field populations, and examine the impact of SVs on parasite phenotypes.

Materials and Methods

Parasite Populations

We examined five laboratory populations of *Schistosoma mansoni* (SmBRE, SmEG, SmLE-PZQ-ER, SmLE-PZQ-ES, SmOR). These parasite populations differ in resistance to anthelmintic drugs, specificity to the snail intermediate host, and were initiated with founder individuals from different geographic regions (Table 1). These parasite populations are maintained by laboratory passage using hamsters as the vertebrate host (where sexual reproduction occurs) and *Biomphalaria* spp. snails as the intermediate host. Two of the populations, SmLE-PZQ-ER and SmLE-PZQ-ES, were derived from the same SmLE stock population <10 years ago (Le Clec'h et al. 2021a). Fresh schistosome worms were obtained from hamster perfusions in the laboratory.

DNA Extraction and Sequencing

We used two different DNA extraction methods for extracting HMW DNA and Nanopore sequencing of schistosome (Table 2).

For the first method, we used the Monarch HMW DNA Extraction Kit for Tissue (NEB #T3060) and manufacturer recommended protocols to isolate DNA from 92 to 152 worms for the SmEG, SmLE-PZQ-ES, and SmOR populations. For the second extraction method we isolated DNA from SmBRE and SmLE-PZQ-ES populations using a developmental version of the Nanobind HMW DNA extraction kit being designed by Circulomics specifically for *C. elegans*. We transferred 100 mg of worms to 2 mL prechilled Protein LoBind microcentrifuge tube. The tube was centrifuged at 16,000×*g* for 15 s. We discarded the resulting supernatant, added 20 μL of Proteinase K and 150 μL of Buffer NL, and resuspended the worm pellet by vortexing 1 to 2 s prior to incubating on a ThermoMixer at 900 rpm and 55°C for 1.5 h. After spinning the tube on a mini centrifuge for 1 s to remove liquid from the cap, we added 20 μL RNase A and incubated on a ThermoMixer at 900 rpm and 55°C for 1 h. After spinning the tube on a mini centrifuge for 1 s to remove liquid from the cap, we added 40 μL Buffer ESB, pulse vortexed 5×, and incubated on ice for 5 min. We centrifuged the tube at 2,000 *g* at RT for 10 min, and transferred 200 μL of the supernatant to a new 1.5-mL Protein LoBind microcentrifuge tube using a wide bore pipette, and discarded the tube containing the precipitate. We added a Nanobind disk to the lysate, followed by 250 μL of isopropanol, and mixed 10 times by inversion, to precipitate DNA and allow binding to the Nanobind disk. The DNA is initially opaque on addition of isopropanol but becomes clear during subsequent washing steps. We mixed the tube on a tube rotator at RT for 15 min. We used the Magnetic rack handling procedure (PacBio PN 102-587-900) for washing the disc with buffers CW1 and CW2. We eluted DNA with the appropriate volume of EB buffer for 7 days at 4°C. We used Qubit dsDNA BR Assay and Nanodrop to record DNA concentration and purity. Nanodrop readings should have A260/A280 in the range of 1.85 to 2.02 and A260/A230 in the range of 2.03 to 2.35. We used the Short Fragment Eliminator kit from PacBio (PN 102-582-400) to eliminate DNA below 25 kb. We stored DNA at 4°C.

Long-Read Sequencing

A single microgram of high-molecular-weight genomic DNA was used for preparing libraries to be sequenced on the MinION. We prepared libraries with either the Genomic DNA by Ligation (SQK-LSK109) kit, or Ligation Sequencing V14 (SQK-LSK114) kit, loaded on flowcells with R9.4.1 or R10.4.1 chemistry. The sequencing runs lasted from 48 to 72 h. Halfway through the run, we paused sequencing, washed the flow cells with EXP-WSH004-XL kit to unclog the pores, loaded more of the same libraries, and resumed sequencing.

Genotyping SVs

Bioinformatic analyses are summarized in Fig. 1. We assessed long-read quantity and N50s with n50 v1.4.2 (Telatin et al. 2021). All reads were quality filtered and trimmed with nanoq v0.10.0 (Steinig and Coin 2022). We trimmed bases with a minimum Phred score of 7 (–min-qual 7), removed the first and last 25 bases (–trim-start 25, –trim-end 25) and any reads that were <500 bp after trimming (–min-len 500). We then mapped the quality-filtered reads to the *S. mansoni* V10 genome assembly (SM_V10) with minimap2 v2.24 (Li 2016) using the "-x map-ont" parameter set. The SM_V10 assembly was accessed via WormBase ParaSite (accession: WBPS19; Harris et al. 2020). Genome coverage was calculated with mosdepth v0.3.2 (Pedersen and Quinlan 2018). Reads were sorted with SAMtools v1.19 (Li et al. 2009) for downstream analyses. SV detection and genotyping occurred in three steps; (i) an initial de novo SV query in each population, followed by (ii) filtering and merging the population SV calls into a single catalogue, then each sample was (iii) re-genotyped using the catalogue of merged variants.

Initial SV Identification

SVs were identified for each individual population with an initial query using cuteSV v2.0.2 (Jiang et al. 2020 2022). We used the following parameters for SV identification: "–max_cluster_bias_INS 100 –diff_ratio_merging_INS 0.3 –max_cluster_bias_DEL 100 –diff_ratio_merging_DEL 0.3 –min_mapq 20 –min_read_len 500 –md 100 –mi 100 –min_support five –max_size 200000 –genotype". The combination of these parameters was chosen to identify larger SVs between 10 bp and 200 kb supported by at least five reads with a minimum mapping quality (MAPQ) of 20. These parameters are recommended for long-read-based SV detection in the cuteSV documentation (https://github.com/tjiangHIT/cuteSV v0.19).

SV Filtering and Merging

After the initial round of SV identification with cuteSV we explored various combinations of QUAL, GQ, and DP filters (discussed in the Results section) before choosing to remove variants with QUAL < 10 and read depth (DP) < 10 using VCFtools v0.1.16 (Danecek et al. 2011). These filters were chosen with the reasoning that DP > 10 would allow us to reasonably calculate allele frequency thresholds in 10% increments and a QUAL > 10 indicates at least 90% confidence in the SV call. The filtered SVs from each population were merged into a single catalogue using SURVIVOR v1.0.7 (Jeffares et al. 2017) and the following parameters: max_distance_between_breakpoints = 100, min_calls = 1, sv_type = 1, sv_strand = 0, dist_estimate = 0, and min_size = 50. These parameters assured that variants

must be present in at least one population and if variants are present in multiple populations they are only merged if their SV boundaries are within 100 bp of one another, the SV is of the same type, and greater than 50 bp.

Final SV Genotyping and Secondary Analyses

The merged catalogue of SVs contains the higher quality SV calls present in one or more *S. mansoni* populations. We provided these variants to SVJedi-graph v1.2.1 (Romain and Lemaitre 2023) for forced variant calling at SVs with a minimum of two supporting reads ('—minsupport 2'). We then converted the population VCF files to tables using the BCFtools v1.14 (Danecek et al. 2011) query. The alternate allele frequency of each variant was calculated by dividing the alternate allele depth (AD) by total depth. Subsequent processing was managed with Pandas v2.0.3.

We used a PCA and Venn diagrams to examine how SV are distributed between the populations. To minimize artifacts caused by biases in the sex distributions between populations, we removed any variants on the Z chromosome, WSR, and mitochondria to focus solely on autosomal SVs. A PCA was calculated for the first two PCs using the PCA() function implemented in scikit-learn v1.2.0. To identify population-specific SVs we used the venn() v0.1.3 package to generate Venn diagrams of the autosomal SVs. Rather than strictly examining presence or absence, we partitioned and categorized each SV as low frequency (<5%) or moderate-high frequency (≥5%). These cutoffs were chosen to minimize false negatives in low coverage populations where a SV may be present at low frequencies but was not sampled in our sequence data. We used the Venn diagram to identify specific variants present at ≥5% in only one population. From this pool, we identified any variants with an allele frequency of ≥95% for downstream analysis. We refer to these SVs as lineage-specific as they are at-or-near fixation (≥95%) in one population but present in <5% in the others populations.

We identified Repeat families de novo using RepeatModeler v2.0.3 (Flynn et al. 2020) to query the entire SM_V10 genome assembly. We used RepeatMasker v4.1.7 (Smit et al. 2015) to quantify repeats in the whole genome, and in the SV alleles separately, using the de novo repeat families as a custom library ("-lib") and the sensitive ("-s") options. We compared the overall repeat content of each category, whole genome vs. SV alleles, using a two-proportion *z*-test. Our null hypothesis was that the overall repeat content of the SV alleles was the same as the rest of the genome. We used RepeatMasker's Smith–Waterman distances between individual insertions and the family consensus sequence as a proxy for the relative age for each insertion. When an element is copied into the genome it is an exact, or near exact replica of the donor element. Over time, elements will acquire independent mutations so the genetic distance between two elements is generally correlated with time. We used a permutation test and 10,000 replicates to compare the Smith–Waterman distances (age) between Gypsy-derived SVs and all Gypsy insertions across the genome.

We used BEDtools v2.31.1 (Quinlan and Hall 2010) to intersect the SV and gene annotations for the SM_V10 genome assembly. Previous work has established six QTLs associated with either praziquantel drug resistance (Le Clec'h et al. 2021a; Chevalier et al. 2024) or cercarial production (Le Clec'h et al. 2021b). We converted these coordinates to the current version of the *S. mansoni* genome by aligning the SM_V7 (GCA_000237925.2) and the SM_V10 assemblies with minimap2 v2.24 (Li 2016) and the "asm5" alignment preset parameter. Coordinates were converted between the assemblies using paftools.js liftover 2.28-r1209 (Li 2016).

Validating SV Calls

We validated a subset of SVs by amplifying the target locus from individual worms in each population and scoring the presence or absence of the variants. For this dataset, we examined 26 population-specific SVs that were between 250 and 750 bp. Primers for each locus were designed using BatchPrimer3 v2.6.1 (You et al. 2008) with an optimum primer size of 20nt, melting temp of 59°C, and size range of 400 to 600 nt. We extracted DNA from individual worms from each population using Qiagen's DNeasy Blood and Tissue kit and performed WGA using Genomiphi V2 DNA Amplification kit from Cytiva. We amplified PCR products using a 25-cycle PCR reaction using TaKaRa Taq polymerase kit and ran 1 µL of the PCR product on a 1% agarose gel (94 V/for 60 min). Gels were visualized on the BIO-RAD ChemiDoc Touch Imaging System and scored as homozygous presence, heterozygous, or homozygous absence for each of the individual worms. We then compared these results to the population allele frequency to determine whether a variant was a true positive, false positive, true negative, or false negative to calculate accuracy, sensitivity, and specificity.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Sandy Smith and John Heaner for providing computational support through Texas Biomedical Research Institute's High-Performance Computing Cluster. Frederic Chevalier, Winka Le Clec'h, Elisha Enabulele, and Kathrin Jutzeler provided parasite material and constructive feedback during the course of this project.

Funding

This research was funded by the National Institute of Allergy and Infectious Diseases (R01 AI166049) and by a Texas Biomedical Forum Grant 20-04866.

Conflict of Interest

The authors declare no competing interests.

Data Availability

We conducted analyses on compute nodes with a maximum of 1 terabyte of memory and 96 cores at the Texas Biomedical Research Institute's high-performance computing cluster. All software and computing environments were managed with Conda v22.9.0. Initial analyses from quality filtering raw sequence data through the final SV genotyping steps were accomplished with a SnakeMake v7.25 (Mölder et al. 2021) workflow. Secondary analyses were conducted in a series of Jupyter v4.2.0 notebooks. All code, including environmental recipe files, notebooks, workflows, and shell scripts are available through GitHub (https://github.com/nealplatt/sch_man_ont) and are permanently accessioned with Zenodo (https://doi.org/10.5281/zenodo.14510995). All sequence data is accessioned under NCBI BioProject PRJNA1206205.

Literature Cited

Agyabenga-Dadzie F, et al. Evaluating the benefits and limits of multiple displacement amplification with whole-genome oxford nanopore sequencing. Mol Ecol. 2025:6:e14094. https://doi.org/10.1111/1755-0998.14094.

Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011:12:363–376. https://doi.org/10.1038/nrg2958.

Alonge M, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell. 2020:182:145–161.e123. https://doi.org/10.1016/j.cell.2020.05.021.

Anderson TJC, LoVerde PT, Le Clec'h W, Chevalier FD. Genetic crosses and linkage mapping in schistosome parasites. Trends Parasitol. 2018:34:982–996. https://doi.org/10.1016/j.pt.2018.08.001.

Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015:6:11. https://doi.org/10.1186/s13100-015-0041-9.

Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 2002:12:1269–1276. https://doi.org/10.1101/gr.88502.

Bellone RR, et al. A de novo 2.3 kb structural variant in MITF explains a novel splashed white phenotype in a thoroughbred family. Anim Genet. 2023:54:752–762. https://doi.org/10.1111/age.13352.

Berriman M, et al. The genome of the blood fluke *Schistosoma mansoni*. Nature. 2009:460:352–358. https://doi.org/10.1038/nature08160.

Brandler WM, et al. Paternally inherited cis-regulatory structural variants are associated with autism. Science. 2018:360:327–331. https://doi.org/10.1126/science.aan2261.

Buddenborg SK, et al. Assembled chromosomes of the blood fluke Schistosoma mansoni provide insight into the evolution of its ZW sex-determination system [preprint]. bioRxiv 456314. https://doi.org/10.1101/2021.08.13.456314

Chakraborty M, et al. Hidden genetic variation shapes the structure of functional elements in Drosophila. Nat Genet. 2018:50:20–25. https://doi.org/10.1038/s41588-017-0010-y.

Cheeseman IH, et al. Population structure shapes copy number variation in malaria parasites. Mol Biol Evol. 2016:33:603–620. https://doi.org/10.1093/molbev/msv282.

Chen X, Dong S, Liu X, Ding N, Xing X. Phenotype of white sika deer due to SCF gene structural variation. Genes (Basel). 2023:14:1035. https://doi.org/10.3390/genes14051035.

Chevalier FD, et al. Oxamniquine resistance alleles are widespread in old world *Schistosoma mansoni* and predate drug deployment. PLoS Pathog. 2019:15:e1007881. https://doi.org/10.1371/journal.ppat.1007881.

Chevalier FD, Le Clec'h W, Berriman M, Anderson TJC. A single locus determines praziquantel response in *Schistosoma mansoni*. Antimicrob Agents Chemother. 2024:68:e0143223. https://doi.org/10.1128/aac.01432-23.

Chiang C, et al. The impact of structural variation on human gene expression. Nat Genet. 2017:49:692–699. https://doi.org/10.1038/ng.3834.

Clément JA, et al. Private selective sweeps identified from next-generation pool-sequencing reveal convergent pathways under selection in two inbred *Schistosoma mansoni* strains. PLoS Negl Trop Dis. 2013:7:e2591. https://doi.org/10.1371/journal.pntd.0002591.

Cretu Stancu M, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun. 2017:8:1326. https://doi.org/10.1038/s41467-017-01343-4.

Czech L, Spence JP, Expósito-Alonso M. Grenedalf: population genetic statistics for the next generation of pool sequencing. Bioinformatics. 2024:40:btae508. https://doi.org/10.1093/bioinformatics/btae508.

Danecek P, et al. The variant call format and VCFtools. Bioinformatics. 2011:27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Delage WJ, Thevenon J, Lemaitre C. Towards a better understanding of the low recall of insertion variants with short-read based variant callers. BMC Genomics. 2020:21:762. https://doi.org/10.1186/s12864-020-07125-5.

Delahaye C, Nicolas J. Sequencing DNA with nanopores: troubles and biases. PLoS One. 2021:16:e0257521. https://doi.org/10.1371/journal.pone.0257521.

Dohm JC, Peters P, Stralis-Pavese N, Himmelbauer H. Benchmarking of long-read correction methods. NAR Genom Bioinform. 2020:2:lqaa037. https://doi.org/10.1093/nargab/lqaa037.

Domínguez M, et al. The impact of transposable elements on tomato diversity. Nat Commun. 2020:11:4058. https://doi.org/10.1038/s41467-020-17874-2.

Dumetz F, et al. Modulation of aneuploidy in leishmania donovani during adaptation to different in vitro and in vivo environments and its impact on gene expression. mBio. 2017:8:e00599-17. https://doi.org/10.1128/mBio.00599-17.

Ebert P, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021:372:eabf7117. https://doi.org/10.1126/science.abf7117.

Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008:9:18. https://doi.org/10.1186/1471-2105-9-18.

Flynn JM, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020:117:9451–9457. https://doi.org/10.1073/pnas.1921046117.

Gong T, Hayes VM, Chan EKF. Detection of somatic structural variants from short-read next-generation sequencing data. Brief Bioinform. 2021:22:bbaa056. https://doi.org/10.1093/bib/bbaa056.

Goubert C, et al. A beginner's guide to manual curation of transposable elements. Mob DNA. 2022:13:7. https://doi.org/10.1186/s13100-021-00259-7.
Guo J, et al. An integrated peach genome structural variation map uncovers genes associated with fruit traits. Genome Biol. 2020:21:258. https://doi.org/10.1186/s13059-020-02169-y.
Guo J, et al. Genome-wide association study reveals 14 new SNPs and confirms two structural variants highly associated with the horned/polled phenotype in goats. BMC Genomics. 2021:22:769. https://doi.org/10.1186/s12864-021-08089-w.
Hackett F. The culture of *Schistosoma mansoni* and production of life cycle stages. Methods Mol Biol. 1993:21:89–99. https://doi.org/10.1385/0-89603-239-6:89.
Harris TW, et al. WormBase: a modern model organism information resource. Nucleic Acids Res. 2020:48:D762–d767. https://doi.org/10.1093/nar/gkz920.
Hassan AH, Haberl B, Hertel J, Haas W. Miracidia of an Egyptian strain of Schistosoma mansoni differentiate between sympatric snail species. The J Parasitol. 2003:89(6):1248–1250. https://doi.org/10.1645/ge-85r.
Hollox EJ, Zuccherato LW, Tucci S. Genome structural variation in human evolution. Trends Genet. 2022:38:45–58. https://doi.org/10.1016/j.tig.2021.06.015.
Huang H, et al. Combining human epigenetics and sleep studies in *Caenorhabditis elegans*: a cross-Species approach for finding conserved genes regulating sleep. Sleep. 2017:40:zsx063. https://doi.org/10.1093/sleep/zsx063.
Hughes JF, Coffin JM. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. Proc Natl Acad Sci U S A. 2004:101:1668–1672. https://doi.org/10.1073/pnas.0307885100.
Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM. Long-read mapping to repetitive reference sequences using Winnowmap2. Nat Methods. 2022:19:705–710. https://doi.org/10.1038/s41592-022-01457-8.
Jeffares DC, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun. 2017:8:14061. https://doi.org/10.1038/ncomms14061.
Jiang T, et al. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. 2020:21:189. https://doi.org/10.1186/s13059-020-02107-y.
Jiang T, Liu S, Cao S, Wang Y. Structural variant detection from long-read sequencing data with cuteSV. Methods Mol Biol. 2022:2493:137–151. https://doi.org/10.1007/978-1-0716-2293-3_9.
Jutzeler KS, et al. Abundant genetic variation is retained in many laboratory schistosome populations [preprint]. bioRxiv 619418. https://doi.org/10.1101/2024.10.21.619418
Kim BY, et al. Single-fly genome assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life. PLoS Biol. 2024:22:e30002697. https://doi.org/10.1371/journal.pbio.3002697
Kosugi S, Terao C. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. Hum Genome Var. 2024:11:18. https://doi.org/10.1038/s41439-024-00276-x.
Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014:15:R84. https://doi.org/10.1186/gb-2014-15-6-r84.
Le Clec'h W, et al. Whole genome amplification and exome sequencing of archived schistosome miracidia. Parasitology. 2018:145:1739–1747. https://doi.org/10.1017/S0031182018000811.
Le Clec'h W, et al. Genetic analysis of praziquantel response in schistosome parasites implicates a transient receptor potential channel. Sci Transl Med. 2021:13:eabj9114. https://doi.org/10.1126/scitranslmed.abj9114.
Le Clec'h W, et al. Genetic architecture of transmission stage production and virulence in schistosome parasites. Virulence. 2021:12:1508–1526. https://doi.org/10.1080/21505594.2021.1932183.
Lee YC, et al. Single-worm long-read sequencing reveals genome diversity in free-living nematodes. Nucleic Acids Res. 2023:51:8035–8047. https://doi.org/10.1093/nar/gkad647.
Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009:25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016:32:2103–2110. https://doi.org/10.1093/bioinformatics/btw152.
Li Y, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020:578:112–121. https://doi.org/10.1038/s41586-019-1913-9.
Liu Y, et al. Repeat and haplotype aware error correction in nanopore sequencing reads with DeChat. Commun Biol. 2024:7:1678. https://doi.org/10.1038/s42003-024-07376-y.
Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020:21:597–614. https://doi.org/10.1038/s41576-020-0236-x.
Lynch M, Bost D, Wilson S, Maruki T, Harrison S. Population-genetic inference from pooled-sequencing data. Genome Biol Evol. 2014:6:1210–1218. https://doi.org/10.1093/gbe/evu085.
Mahmoud M, et al. Structural variant calling: the long and the short of it. Genome Biol. 2019:20:246–246. https://doi.org/10.1186/s13059-019-1828-7.
Mölder F, et al. Sustainable data analysis with snakemake. F1000Res. 2021:10:33. https://doi.org/10.12688/f1000research.29032.2.
Moore DV, Sandground JH. The relative egg producing capacity of *Schistosoma mansoni* and *Schistosoma japonicum*. Am J Trop Med Hyg. 1956:5:831–840. https://doi.org/10.4269/ajtmh.1956.5.831.
Nair S, et al. Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. Mol Biol Evol. 2007:24:562–573. https://doi.org/10.1093/molbev/msl185.
Nair S, et al. Adaptive copy number evolution in malaria parasites. PLoS Genet. 2008:4:e1000243. https://doi.org/10.1371/journal.pgen.1000243.
Nelwan ML. Schistosomiasis: life cycle, diagnosis, and control. Curr Ther Res Clin Exp. 2019:91:5–9. https://doi.org/10.1016/j.curtheres.2019.06.001.
Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2018:176:1410–1422. https://doi.org/10.1104/pp.17.01310.
Pang AW, et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 2010:11:R52. https://doi.org/10.1186/gb-2010-11-5-r52.
Park SK, et al. Mechanism of praziquantel action at a parasitic flatworm ion channel. Sci Transl Med. 2021:13:eabj5832. https://doi.org/10.1126/scitranslmed.abj5832.
Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018:34:867–868. https://doi.org/10.1093/bioinformatics/btx699.
Platt RN 2nd, et al. Genomic analysis of a parasite invasion: colonization of the Smericas by the blood fluke *Schistosoma mansoni*. Mol Ecol. 2022:31:2242–2263. https://doi.org/10.1111/mec.16395.
Platt RN 2nd, Blanco-Berdugo L, Ray DA. Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biol Evol. 2016:8:403–410. https://doi.org/10.1093/gbe/evw009.
Pollo SMJ, et al. Benchmarking hybrid assemblies of Giardia and prediction of widespread intra-isolate structural variation. Parasit

Vectors. 2020:13:108. https://doi.org/10.1186/s13071-020-3968-8.
Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005:21 Suppl 1:i351–i358. https://doi.org/10.1093/bioinformatics/bti1018.
Protasio AV, et al. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. PLoS Negl Trop Dis. 2012:6:e1455. https://doi.org/10.1371/journal.pntd.0001455.
Quah FX, et al. A pangenomic perspective of the Lake Malawi cichlid radiation reveals extensive structural variation driven by transposable elements [preprint]. bioRxiv 587230. https://doi.org/10.1101/2024.03.28.587230
Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010:26:841–842. https://doi.org/10.1093/bioinformatics/btq033.
Ravenhall M, et al. An analysis of large structural variation in global *Plasmodium falciparum* isolates identifies a novel duplication of the chloroquine resistance associated gene. Sci Rep. 2019:9:8287. https://doi.org/10.1038/s41598-019-44599-0.
Reis ALM, et al. The landscape of genomic structural variation in indigenous Australians. Nature. 2023:624:602–610. https://doi.org/10.1038/s41586-023-06842-7.
Romain S, Lemaitre C. SVJedi-graph: improving the genotyping of close and overlapping structural variants with long reads using a variation graph. Bioinformatics. 2023:39:i270–i278. https://doi.org/10.1093/bioinformatics/btad237.
Ruggieri AA, et al. A butterfly pan-genome reveals that a large amount of structural variation underlies the evolution of chromatin accessibility. Genome Res. 2022:32:1862–1875. https://doi.org/10.1101/gr.276839.122.
Sabina J, Leamon JH. Bias in whole genome amplification: causes and considerations. Methods Mol Biol. 2015:1347:15–41. https://doi.org/10.1007/978-1-4939-2990-0_2.
Sarkozy P, Jobbágy Á, Antal P. Calling homopolymer stretches from raw nanopore reads by analyzing k-mer dwell times. In European Medical and biological engineering confernce; Singapore. Springer. 2017:241–244.
Schlötterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. Nat Rev Genet. 2014:15:749–763. https://doi.org/10.1038/nrg3803.
Sibbesen JA, Maretty L, Krogh A. Accurate genotyping across variant classes and lengths using variant graphs. Nat Genet. 2018:50:1054–1059. https://doi.org/10.1038/s41588-018-0145-5.
Smeds L, Huson LSA, Ellegren H. Structural genomic variation in the inbred scandinavian wolf population contributes to the realized genetic load but is positively affected by immigration. Evol Appl. 2024:17:e13652. https://doi.org/10.1111/eva.13652.
Smit AF, Hubley R, Green P. RepeatMasker Open-4.0 2015. https://www.repeatmasker.org/faq.html#faq3.
Steinig E, Coin L. Nanoq: ultra-fast quality control for nanopore reads. J Open Source Softw. 2022:7:2991. https://doi.org/10.21105/joss.02991.
Steinmann P, Keiser J, Bos R, Tanner M, Utzinger J. Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. Lancet Infect Dis. 2006:6:411–425. https://doi.org/10.1016/S1473-3099(06)70521-7.
Sudmant PH, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015:526:75–81. https://doi.org/10.1038/nature15394.
Telatin A, Fariselli P, Birolo G. Seqfu: a suite of utilities for the robust and reproducible manipulation of sequence files. Bioengineering (Basel). 2021:8:59. https://doi.org/10.3390/bioengineering8050059.
Thoendel M, et al. Impact of contaminating DNA in whole-genome amplification kits used for metagenomic shotgun sequencing for infection diagnosis. J Clin Microbiol. 2017:55:1789–1801. https://doi.org/10.1128/JCM.02402-16.
Trost B, Loureiro LO, Scherer SW. Discovery of genomic variation across a generation. Hum Mol Genet. 2021:30:R174–r186. https://doi.org/10.1093/hmg/ddab209.
Valentim CL, et al. Genetic and molecular basis of drug resistance and species-specific drug action in schistosome parasites. Science. 2013:342(6164):1385–1389. https://doi.org/10.1126/science.1243106.
Van't Hof AE, et al. The industrial melanism mutation in British peppered moths is a transposable element. Nature. 2016:534:102–105. https://doi.org/10.1038/nature17951.
Vianney TJ, et al. Genome-wide analysis of *Schistosoma mansoni* reveals limited population structure and possible praziquantel drug selection pressure within Ugandan hot-spot communities. PLoS Negl Trop Dis. 2022:16:e0010188. https://doi.org/10.1371/journal.pntd.0010188.
Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013:14:125–138. https://doi.org/10.1038/nrg3373.
Wendt G, et al. A single-cell RNA-seq atlas of *Schistosoma mansoni* identifies a key regulator of blood feeding. Science. 2020:369:1644–1649. https://doi.org/10.1126/science.abb7709.
Yang L. A practical guide for structural variation detection in the human genome. Curr Protoc Hum Genet. 2020:107:e103. https://doi.org/10.1002/cphg.103.
Ye F, et al. Characteristics and filtering of low-frequency artificial short deletion variations based on nanopore sequencing. Gigascience. 2025:14:giaf018. https://doi.org/10.1093/gigascience/giaf018.
You FM, et al. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinformatics. 2008:9:253. https://doi.org/10.1186/1471-2105-9-253.
Young ND, et al. Whole-genome sequence of *Schistosoma haematobium*. Nat Genet. 2012:44:221–225. https://doi.org/10.1038/ng.1065.Zhang K, et al. Genomic structural variation contributes to phenotypic change of industrial bioethanol yeast Saccharomyces cerevisiae. FEMS Yeast Res. 2016:16:fov118. https://doi.org/10.1093/femsyr/fov118.
Zong W, et al. Associations of genome-wide structural variations with phenotypic differences in cross-bred Eurasian pigs. J Anim Sci Biotechnol. 2023:14:136. https://doi.org/10.1186/s40104-023-00929-x.

Associate editor: John Wang