Whole genome amplification and exome sequencing of archived schistosome miracidia

Winka Le Clec'h1, Frédéric D. Chevalier1, Marina McDew-White1, Fiona Allan2, Bonnie L. Webster2, Anouk N. Gouvras2, Safari Kinunghi3, Louis-Albert Tchuem Tchuente4,5, Amadou Garba6, Khalfah A. Mohammed7, Shaali M. Ame8, Joanne P. Webster9, David Rollinson2, Aidan M. Emery2 and Timothy J. C. Anderson1

1Department of Genetics, Texas Biomedical Research Institute, PO Box 760549, San Antonio, TX 78245-0549, USA; 2Department of Life Sciences, The Natural History Museum, Cromwell Road, London SW7 5BD, UK; 3National Institute for Medical Research, Mwanza Research Centre, Mwanza, United Republic of Tanzania; 4Laboratoire de Parasitologie et Ecologie, Université de Yaoundé 1, Yaoundé, Cameroon; 5Center for Schistosomiasis & Parasitology, P.O. Box 7244, Yaoundé, Cameroon; 6Réseau International Schistosomoses, Environnement, Aménagement et Lutte (RISEAL-Niger), 333, Avenue des Zarmakoye, B.P. 13724, Niamey, Niger; 7Ministry of Health, Helminth Control Laboratory Unguja, Zanzibar, United Republic of Tanzania; 8Public Health Laboratory – Ivo de Carneri, Pemba, United Republic of Tanzania and 9Department of Pathobiology and Population Sciences, Centre for Emerging, Endemic and Exotic Diseases, Royal Veterinary College, University of London, AL9 7TA, UK

Abstract

Adult schistosomes live in the blood vessels and cannot easily be sampled from humans, so archived miracidia larvae hatched from eggs expelled in feces or urine are commonly used for population genetic studies. Large collections of archived miracidia on FTA cards are now available through the Schistosomiasis Collection at the Natural History Museum (SCAN). Here we describe protocols for whole genome amplification of Schistosoma mansoni and Schistosoma haematobium miracidia from these cards, as well as real time PCR quantification of amplified schistosome DNA. We used microgram quantities of DNA obtained for exome capture and sequencing of single miracidia, generating dense polymorphism data across the exome. These methods will facilitate the transition from population genetics, using limited numbers of markers to population genomics using genome-wide marker information, maximising the value of collections such as SCAN.

Introduction

We currently lack tools for effective genome-wide characterization of schistosomes from human populations for two main reasons. First, there is the practical problem of obtaining suitable material for genomic analyses: schistosome adults live in the blood vessels and only eggs expelled in feces (Schistosoma mansoni and japonicum) or urine (Schistosoma haematobium) are available from patients. Microscopic miracidium larvae (70 × 140 µm) hatched from these eggs can be used to infect snails and the cercariae larvae produced can be used to infect laboratory rodents to recover adult worms, but this approach is cumbersome, ethically questionable and imposes strong selection and potential bias (Gower et al., 2007). An alternative approach is to hatch these larvae and preserve them individually on FTA cards for genetic analysis (Gower et al., 2007). This approach has been widely used and has greatly improved our understanding of many aspects of schistosome population biology (Steinauer et al., 2010; Webster et al., 2012; Gower et al., 2013, 2017). However, these studies can be severely constrained by the minute amount of DNA and therefore the relatively small number of genetic markers that can be successfully multiplexed, PCR amplified and analysed from a single miracidium. Typically these studies utilize just 8–20 microsatellite markers (Gower et al., 2011, 2013, 2017; Glenn et al., 2013; Steinauer et al., 2013; Aemero et al., 2015; Webster et al., 2015) or 1–3 gene regions (Van den Broeck et al., 2015; Li et al., 2017). The use of FTA cards has allowed the creation of large archived collections of FTA-preserved miracidia maintained at the Schistosomiasis Collection at the Natural History Museum (SCAN) (Emery et al., 2012). SCAN currently houses around 300,000 miracidia collected from over 7000 people from 14 countries, sampled within the past decade, providing valuable geographical and temporal populations of samples for molecular analysis.

The size and complex architecture of schistosome genomes provide an additional obstacle. Whole genome sequences of the three principal schistosome species infecting humans (S. mansoni, S. haematobium and S. japonicum) are now available, but these are large (363–400 Mb), and riddled with transposable elements and repeats, which comprise 40–50% of the genome and make alignment problematic (Berriman et al., 2009; Zhou et al., 2009; Protasio et al., 2012; Young et al., 2012). The large genome size also makes whole genome sequencing from populations prohibitively expensive, even with falling sequencing costs. Furthermore, the coding regions of the genome (the exome) that are of primary interest for
many analyses measure ~15 Mb and constitute just 4% of the total genome (Young et al., 2012).

Exome capture methods utilize RNA baits that hybridize with the targeted exome sequences, so they can be isolated from non-coding DNA. This approach has been widely used to rapidly and economically sequence thousands of human exomes and identify coding variants associated with disease (Rabbani et al., 2014; Pehlivan et al., 2014; Shaw et al., 2017). There are several attractive features of exome sequencing for schistosome miracidia: (i) this approach is scalable to working with hundreds of individual miracidia, (ii) repetitive regions of the genome that are problematic to align are removed, (iii) contaminating sequences amplified from FTA cards (e.g. bacteria from fecal matter and water) are eliminated, and (iv) reducing the genome size from 363 to 15 Mb allows sequencing to high read depth, providing robust scoring of variable sites. We have previously described an exome capture approach that we used for whole genome amplified DNA prepared from adult S. mansoni (Chevalier et al., 2014). Here we adapt this approach for single miracidia. We describe methods for (i) whole genome amplification (WGA) of miracidia DNA directly from FTA cards, (ii) qPCR methods for quantifying of schistosome DNA amplified and (iii) exome capture, sample multiplexing and sequencing of amplified DNA. We apply these methods to a subset of miracidia from SCAN to demonstrate the utility of these methods. Furthermore, we describe protocols for both S. mansoni and S. haematobium, the two species responsible for more than 99% of human infections (Hotez et al., 2006).

Materials and methods

Ethics statement

EU-CONTRAST specimens were collected in line with project ethical approval granted by ethical committees of the Ministry of Health Dakar, Senegal, the Niger National Ethical Committee, Comité National d’Ethique, Cameroon. For specimens from the Schistosomiasis Consortium for Operational Research and Evaluation (SCORE), ethical approvals were granted by Royal Veterinary College 2015 1327; Imperial College Research Ethics Committee and SCI Ethical approval (EC no: 03.36. R&D no: 03/SB/033E); the National Institute for Medical Research (NIMR, reference no. NIMR/HR/R.8a/Vol. IX/1022); Zanzibar Medical Research Ethics Committee in Stonetown, Zanzibar (ZAMREC, reference no. ZAMREC 0003/Sept/011); University of Georgia Institutional Review Boards, Athens, GA (2011-10353-1); Niger Republic National Consulate (reference no. 012/2010/CCNE).

All local requirements were met, including obtaining written informed consent from adults (including parents/legal guardians of children included in the studies) and followed by obtaining the assent from children included. Following sampling, a praziquantel treatment (40 mg kg⁻¹) was offered to infected participants.

Field collection sites

Schistosoma mansoni miracidia were collected as part of EU-CONTRAST activities from individual patients in three different West African countries in 2007: 27 patients from two locations in Senegal, 17 patients from two locations in Niger and six patients from one location in Cameroon (Supplementary File 1) (Gower et al., 2011; Webster et al., 2013). Additionally, S. mansoni miracidia were collected as part of the SCORE programme from 64 children in seven villages on the shores of Lake Victoria in Tanzania (East Africa) in January 2012 (Ezeamama et al., 2016) (Supplementary File 1).

Schistosoma haematobium miracidia were collected through the SCORE Zanzibar Elimination of Schistosomiasis Transmission (ZEST) activities in 2011 from 80 children in 26 locations of the Zanzibar archipelago (East Africa) and in 2013 from 57 patients in 10 locations of Niger (West Africa) (Supplementary File 1).

All samples were transferred to SCAN for checking, curation and storage either immediately (SCORE), or at the close of the project (EU-CONTRAST).

Collection of miracidia

Our methods for recovery of S. mansoni eggs followed Visser and Pitchford (1972) and Gower et al. (2013) with some minor differences. In brief, (i) each S. mansoni infected stool sample (approximately 2 g) was homogenized with a plastic spatula through a 212 µm wire mesh sieve (Endecotts Ltd, London, UK) and washed through with approximately 1 L of locally available bottled water, (ii) the tray contents were transferred to the inner nylon mesh (200 µm pore size) of a Pitchford funnel assembly (Visser and Pitchford, 1972), washed with additional water (up to an additional 1 L), and the inner mesh removed from the assembly, (iii) the washed, filtered homogenate was drained into a 100 × 15 mm Petri dish (BD Biosciences, Erembodegem, Belgium) by opening the tap at the base of the Pitchford funnel. The Petri dish containing the homogenate was then left in bright ambient light (not direct sunlight) to allow hatching of miracidia. Schistosoma haematobium eggs were concentrated from each infected urine by sedimentation, then rinsed in bottled mineral water before transfer into a clean Petri dish containing mineral water and exposed to light to facilitate hatching (Webster et al., 2012).

Both miracidia from S. mansoni and S. haematobium were captured individually, under a binocular microscope, in 3 µL of water and spotted onto an indicating Whatman FTA Classic Indicating card (GE Healthcare Life Sciences, Amersham, UK) using a 20 µL micropipette. Spotted samples (up to approximately 80 per card) can be easily located on the cards because the pink dye turns white after water contact. The cards were then allowed to dry for 1 h at room temperature before being stored in a sealed plastic bag and then shipped to UK, ultimately to be stored in the SCAN repository.

Preparation of FTA samples for WGA

For each sample prepared, a 2 mm diameter disc was removed from the centre of the dye-cleared area using a 2 mm Harris Micro-punch (GE Healthcare Life Sciences, Amersham, UK). This 2 mm disc corresponds to the entire spot containing the whole miracidium. Each punch was placed individually in 500 µL Matrix screw cap two-dimensional barcode storage tubes (Thermo Fisher Scientific, Hemel Hempstead, UK) and shipped to Texas Biomedical Research Institute for further preparation.

The punches were individually transferred into 1.5 mL sterile microtubes using sterile tips. Punches were washed three times with FTA Purification Reagent (GE Healthcare Life Sciences, Logan, Utah, USA) then rinsed two times with TE⁻¹ buffer (10 mM Tris, 0.1 mM EDTA, pH 8). Washing and rinsing steps were performed by adding 200 µL of solution in each tube followed by 5 min of incubation on a nutating mixer (24 RPM) at room temperature and then discarding the solution while minimizing contact between the pipette tip and the punch. Punches were finally dried in tubes during 10 min at 56 °C in a dry bath incubator (Chevalier et al., 2016).
**Whole genome amplification**

We conducted WGA on 1–4 miracidia from each patient from all the sites sampled until a positive WGA was obtained. We performed WGA on each punch using the Illuṣtra GenomiPhi V2 DNA Amplification kit (GE Healthcare Life Sciences, Logan, Utah, USA). Punches were transferred into 0.2 mL sterile tubes using a sterile tip. Reactions were performed following the manufacturer’s instructions, immersing each punch in 9 µL of sample buffer and keeping tubes on ice after the denaturation step (3 min at 95 °C) while adding a mix of 9 µL of reaction buffer and 1 µL of Phi29 polymerase. After the amplification step (2 h 30 min at 30 °C and 10 min at 65 °C), we added 130 µL of sterile water into the reaction tube and purified the 150 µL of amplified genome with the SigmaSpin™ Sequencing Reaction Clean-up (Sigma-Aldrich, Laramie, Wyoming, USA), following the manufacturer’s protocol. We quantified purified samples using the Qubit dsDNA BR assay (Invitrogen, Grand Island, New York, USA).

**Real-time quantitative PCR on schistosome WGA samples**

We performed real-time quantitative PCR (qPCR) reactions to estimate the proportion of schistosome genome in each WGA sample. This was done because FTA samples can contain contaminant DNA, such as human or bacterial DNA, which is co-amplified with schistosome DNA during the WGA step. In this assay, we amplified the *S. mansoni* α-tubulin 1 (sat-1) gene, which is present in low copy (gene number: Smp_090120; accession number: M80214) (Webster et al., 1992) or the putative *S. haematobium* α-tubulin 2. The latter was identified by performing a blastn (v2.2.9) using the *S. mansoni* α-tubulin (Duvaux-Miret et al., 1991; accession number: S79195) against the *S. haematobium* reference genome (SchHae_1.0; assembly accession number: GCA_000699445.1). Reactions were performed in duplicate using the ABI prism 7900HT Sequence Detection system (Applied Biosystems, Carlsbad, California, USA) as follows: 95 °C for 10 min, then 40 cycles of 95 °C for 15 s and 60 °C for 1 min. Duplicate reactions showing a difference in Ct greater than one were rerun. We examined the melting curve (60–95 °C) at the end of each assay to verify the uniqueness of the PCR products generated. The reaction mixture consisted of 5 µL of SYBR Green MasterMix (Applied Biosystems, Carlsbad, California, USA), 0.3 µL of 10 µM forward primer (*S. mansoni*: CGAAATTTGAG TTTGCTGTGT; *S. haematobium*: GGTGGTACTGGTTCTGG TTT) and 10 µM reverse primer (*S. mansoni*: TGTAGGTTGGAC GCCTCTATTAC; *S. haematobium*: AAAGCACAATCCGAATG TTCTAA) amplifying 229 bp of the sat-1 gene for *S. mansoni* and 178 bp of the α-tubulin 2 for *S. haematobium*, 3.4 µL of sterile water and 1 µL of total DNA template (normalized at 20 ng·µL⁻¹). We plotted standard curves using seven dilutions of a purified sat-1 PCR product for *S. mansoni* (sat-1 copies/µL⁻¹: 2.19 × 10³, 2.19 × 10⁴, 2.19 × 10⁵, 2.19 × 10⁶, 2.19 × 10⁷, 2.19 × 10⁸) or seven dilutions of a purified α-tubulin 2 PCR product for *S. haematobium* (α-tubulin 2 copies/µL⁻¹: 1.29 × 10¹, 1.29 × 10², 1.29 × 10³, 1.29 × 10⁴, 1.29 × 10⁵, 1.29 × 10⁶, 1.29 × 10⁷). The number of sat-1 or α-tubulin 2 copies in each sample was estimated according to the standard curve.

**Exome library preparation and sequencing**

We captured schistosome (*S. mansoni* or *S. haematobium*) exomes using the SureSelectXT2 Target Enrichment System (Agilent) according to the manufacturer’s protocol. For each library (i.e. each sample), we sheared 1 µg of WGA DNA by adaptive focused acoustics (Duty factor: 10%; Peak Incident Power: 175; Cycles per Burst: 200; Duration: 180 s) in AFA tubes, using a Covaris S220 instrument with SonoLab software version 7 (Covaris, Inc., Woburn, Massachusetts, USA), to recover fragmented DNA between 150 and 200 bp. We used five PCR cycles for the pre-capture and eight PCR cycles for post-capture amplifications. In order to make the capture step cost-efficient, we added unique indices to each library prior to exome capture (pre-capture indexing method), pooled equal amounts of DNA from 16 libraries and then performed the capture. A capture requiring a total of 750 ng indexed DNA, we pooled 46.875 ng of each of the 16 libraries. To minimize uneven capture of each samples, we pooled libraries from samples showing similar quantities of schistosome genomes previously estimated by qPCR. The capture was performed using sets of baits from the SureSelectXT Target Enrichment System (Agilent, Lexington, Massachusetts, USA) (120 bp RNA molecules) specific to each species.

The design of the baits used to capture the *S. mansoni* exome (SureSelect design ID: S0398493) has already been described in Chevalier et al. (2014). We use our experience with the *S. mansoni* bait design to improve our design strategy for *S. haematobium*. Unlike the *S. mansoni* design, (i) we included all exons that passed the low complexity threshold (rather than only the exons to which sequences can be unambiguously aligned); (ii) we included exons from vaccine candidates (Supplementary File 2); and (iii) we designed baits to ensure effective capture of SHUTL1 OR gene (Sha_104171), a drug target of particular interest (Taylor et al., 2017). The design (SureSelect design ID: S0742423) was made using the latest *S. haematobium* genome version (SchHae_1.0; assembly accession number: GCA_000699445.1) and its corresponding annotation. The final set of *S. haematobium* baits included 156,004 from the nuclear genome and 67 from the mitochondrial genome. These cover 96% (62,106/64,642) of the exons and account for 94% of the exome length (15,002,706 bp/15,895,612 bp). The sequences covered by baits are referred to as the bait regions. Each captured exon was covered by 2.59 baits on average.

We sequenced the pooled barcoded exome libraries using 100 bp pair-end reads (16 samples per lane of flowcell) on HiSeq 2500 sequencer (Illumina). Raw sequence data have been submitted to the NCBI Sequence Read Archive under accession numbers SRP136210 and SRP136277.

**Sequencing data analyses**

We aligned the sequencing data against the *S. mansoni* reference genome (v5; ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/genome/Assembly-v5/ARCHIVE/sma_v5.0.chr.fa.gz) or the *S. haematobium* reference genome (SchHae_1.0; assembly accession number: GCA_000699445.1) using BWA (v0.7.12) (Li and Durbin, 2009) and SAMtools (v1.2) (Li et al., 2009). Realignment around indels was performed using GATK (v3.3-0-g37228af) (McKenna et al., 2010; DePristo et al., 2011). PCR duplicates were marked using picard (v1.136). Q-score recalibration and variant (SNP/indel) calling by the UnifiedGenotyper module were performed using GATK. Bait representation, capture efficiency and read depth analyses were performed using BEDTools (v2.21.0) (Quinlan and Hall, 2010). The variant calling set was filtered using VCFtools (v0.1.14) ( Danecek et al., 2011) to retain only variants in the bait regions and with a minimum read depth of 10 and a minimum genotype quality score of 80.

**Statistics and data representation**

Statistics analyses, calculation of capture efficiency, calculation of the read depth ratio of sex chromosome, analysis of read depth surrounding bait regions and data visualizations were performed using R (v3.3.1). Figure 1 summarizes our pipeline for generating and analysing exome sequence from single miracidia.
**Results**

**WGA of single miracidia**

We performed WGA on a total of 412 FTA-preserved samples. These include 101 *S. mansoni* samples from East Africa (Tanzania) and 90 *S. mansoni* samples from West Africa (Senegal, Niger and Cameroon) (1–3 samples per patient) as well as 138 *S. haematobium* samples from East Africa (Zanzibar archipelago) and 83 FTA preserved *S. haematobium* samples from West Africa (Niger) (1–4 samples per patient). This resulted in an average of 3.8 ± 0.66 µg of DNA (mean ± S.D.) from each *S. mansoni* sample and an average of 2.8 ± 0.48 µg of DNA from each *S. haematobium* sample.

The amount of schistosome DNA per sample is critical for downstream genomic assays: samples with no schistosome DNA need to be removed, while samples with low levels (<100 copies) of schistosome template DNA result in failed exome capture. To estimate the quantity of schistosome DNA in each WGA, we quantified a low copy gene (*sat-1* for *S. mansoni* and *α*-tubulin 2 for *S. haematobium*) using qPCR. While DNA was successfully amplified from almost 100% of the spots (Table 1), the qPCR assay revealed that only 59.4 and 41.11% of the whole genome amplified samples were positive for *S. mansoni* from the East and West African samples, respectively. For the *S. haematobium* samples, 47.10% of schistosome-positive WGA samples were from the East and 66.26% from the West African samples. We found no *S. mansoni*-positive samples among 14 FTA punches examined from Cameroon (Table 1).

The quantity of *S. mansoni* DNA obtained from single miracidia after WGA varied over five orders of magnitude (Fig. 2) among the 97 *S. mansoni*-positive samples. 47.10% of schistosome-positive WGA samples were from the East and 66.26% from the West African samples. We observed similar heterogeneity in schistosome DNA quantity for WGA from the 65 *S. haematobium*-positive samples from...
### Table 1. Whole genome amplification and qPCR assay summary statistics

<table>
<thead>
<tr>
<th>Location</th>
<th>Schistosome spp.</th>
<th>Collection date</th>
<th>No. of patients with schistosome-positive WGAs/total no. WGAs performed</th>
<th>No. of patients with schistosome-positive WGA/no. of patients tested</th>
<th>Average no. of spots tested per patient</th>
<th>Average (+SD) DNA concentration recovered after WGA (ng µL⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>East Africa</td>
<td>S. mansoni</td>
<td>2006</td>
<td>7</td>
<td>13/28</td>
<td>46.42%</td>
<td>24.63 ± 5.04</td>
</tr>
<tr>
<td></td>
<td>S. haematobium</td>
<td>2013</td>
<td>55</td>
<td>83</td>
<td>66.26%</td>
<td>21.21 ± 2.85</td>
</tr>
<tr>
<td>W. Africa</td>
<td>S. mansoni</td>
<td>2011</td>
<td>65</td>
<td>138</td>
<td>47.10%</td>
<td>22.45 ± 4.65</td>
</tr>
<tr>
<td></td>
<td>S. haematobium</td>
<td>2011</td>
<td>62</td>
<td>80</td>
<td>47.10%</td>
<td>22.22</td>
</tr>
</tbody>
</table>

Summary of the number and percentage of positive schistosome samples, the number of patients from which schistosome-positive WGAs were obtained, the average number of spots tested per patient in order to get at least one positive for schistosome, and the average (±SD) DNA concentration recovered after WGA in ng µL⁻¹. We initially conducted WGA from a single miracidium per patient; further WGA were done only if the initial sample was schistosome-negative. Only shown for schistosome-positive samples. (Table 2, Fig. 3).

We show exome sequencing data from 8 S. mansoni miracidia from a single Tanzanian village (Kigongo, Supplementary File 1) and from eight S. haematobium miracidia from three villages on Pemba island (Zanzibar archipelago: Chambani, Ngwachani and Chanjamjamiri; Supplementary File 1). We observed efficient exome capture and sequencing for both species (Table 2). Sequencing of the exome capture libraries generated from six to 26 million reads for S. mansoni and from five to 18 million reads for S. haematobium (Table 2). A very high proportion of these reads were mapped to their respective genomes (>95% on average, respectively) with only one sample from each species showing <90% of mapped reads (84 and 89%, respectively). The remaining DNA sequences, which were not mappable, could either be highly divergent schistosome sequences or more likely human or microbial contaminants present in the WGA sample and captured with the schistosome DNA.

We captured more than 99% of the bait regions and this was consistent across all eight libraries of each species (Table 2). We obtained an average of 51.74 mean (30.62 median) read depth in the bait regions for the S. mansoni libraries and an average of 38.07 mean (30.62 median) read depth for the S. haematobium libraries (Table 2). Read depth was even across the exome (Fig. 3), except in parts of the Z chromosome in S. mansoni females where read depth is halved (females are ZZ). These genome regions (position 3.5–19.5 and 23.5–31 Mb for S. mansoni (Lepesant et al., 2012)) correspond to the non-recombining heterochromatin domain of the W chromosome. The read depth in the Z-linked region allows us to determine the parasite sex in silico using a read depth ratio between the Z-linked region and the rest of the chromosome. Among the samples present in Table 2, we obtained an equal sex ratio. The S. haematobium genome assembly is poorly resolved and the Z-linked region still remains undefined (Table 2, Fig. 3).

We also captured genome regions outside the bait regions, as shown previously (Chevalier et al., 2014). The read depth declines with distance from the bait regions as expected (Table 2). When we included data up to 250 bp around the bait regions, the mean and the median read depth reached ~72 and ~37% of the initial bait regions, respectively. These observations were similar for both S. mansoni and S. haematobium. Therefore, our exome capture provides adequate sequence read depths for at least 250 bp surrounding bait regions allowing us to obtain information regarding adjacent genome regions containing promoters, transcription binding sites and other features of interest.

We were able to robustly identify variable sites in the exome capture libraries from each species (Table 3). The raw variant calling was in silico effective. Among them, the vast majority (96%)
we were biallelic sites, <1% was multiallelic sites and between 3 and 4% were identified as indel sites.

**Discussion**

**Utility of WGA and qPCR quality check for FTA-preserved miracidia**

WGA provides a very effective approach to generating large amounts of DNA from single miracidia for genome wide analyses (Valentim et al., 2009; Shortt et al., 2017). This approach has previously been used to generate micrograms of DNA from a S. mansoni laboratory genetic cross with a minimal error rate (0.45%) induced by the WGA step (Valentim et al., 2009). The current study extends this approach to WGA of archived miracidia collected in the field. Such material is more challenging compared with laboratory prepared miracidia because the miracidia on FTA cards may contain extensive environmental contamination, such as fecal bacteria or human DNA. Two features of the approach used are worth emphasising. First, the WGA method uses 2 mm FTA punches directly immersed in the WGA reaction, eliminating the need for initial DNA preparation. This produced around 3 µg of material that is sufficient for next generation sequencing applications that require at least 500 ng to 1 µg of DNA. Second, we developed a simple qPCR assay for quantifying amounts of S. mansoni and S. haematobium DNA present in the WGA material. This allows identification of schistosome-positive samples and quantification of schistosome DNA, which is a critical step for the identification of WGA samples suitable for library preparation and for grouping samples in pools with similar amounts of schistosome DNA to minimize variation in read depth.

We observed variation in the numbers of positive schistosome HGAs, ranging from 46.42 to 66.26% for S. mansoni or S. haematobium populations, respectively. These results most likely reflect variations in the miracidia collection protocol (variation in egg washing, in spot sizes on FTA cards) and/or in the training of people collecting miracidia. We observed higher and more uniform concentration of S. haematobium DNA in WGA material, when compared with S. mansoni DNA (Fig. 2). The higher variation in S. mansoni DNA amount is consistent with fecal contamination, as less contamination is expected in S. haematobium isolated from urine. We recommend that future miracidia collections, specifically for genomic work, should wash S. mansoni miracidia in clean water, prior to pipetting onto FTA cards, minimizing contamination.

Among schistosome-positive samples, we removed those showing low quantity of schistosome DNA (<100 copies of tubulin gene after WGA) because only higher schistosome DNA quantities gave good sequencing data.

**Efficient exome capture from FTA preserved miracidia**

Preparation of exome capture libraries using WGA material from either S. mansoni or S. haematobium led to high read depth exome sequencing data despite five orders of magnitude variation in the quantity of schistosome DNA present in the WGAs. Using very conservative calling parameters, we scored 85 133 variable sites in S. mansoni and 60 193 in S. haematobium from just eight individual miracidia of each species, which is equivalent to one variant every 166 bp for S. mansoni and every 249 bp for S. haematobium in the bait regions sequenced. This provides a rich source of variation for population exomic analyses and will be the focus of future studies with large numbers of samples.

Our exome capture method has several useful characteristics for working with field-collected miracidia. Exome capture effectively pulls out schistosome DNA (exons) from contaminating human or bacterial DNA so we can minimize the amount of contaminant sequence generated. Capture is very efficient and consistent (more than 99% of the expected region is captured for each sample) so we can sequence to high read depth to generate robust variant calls, or use read depth to estimate copy number. Here, for example, we can use read depth information from the S. mansoni Z-linked sex chromosomes for in silico sexing. The main disadvantages of exome sequencing are some highly variable genes may be poorly captured and many non-coding regions important for gene regulation are not captured (Biesecker et al., 2011), although we do effectively capture non-coding regions adjacent to bait sequences. We note that whole genome sequencing can also be done using WGA miracidia, if precautions are made to minimize contaminations during collection of miracidia.

**Costs of exome sequencing miracidia**

Exome capture methods are more expensive than other reduced representation library preparation approaches, such as RADseq (Harvey et al., 2016). However, exome capture targets specific genome regions of interest and in a consistent manner for each sample, and shows lower levels of genotyping error and allelic drop out (Attard et al., 2017) than RADseq and related methods, so it has several advantages. The baits are by far the most expensive part of this method (~$US900 for a given capture) but the pre-capture indexing method allows pooling of up to 16 samples prior the capture. This brings the capture cost down to $56.25 per sample, while the other reagents required for library preparation are $83.75 per sample. The overall cost per sample is therefore approximately $140. The cost of a flow cell lane for a HiSeq 2500, sufficient to sequence 16 libraries, is currently around $2000 in our facility, so the final cost is $265 per sample for an average read depth of 50. To achieve the same read depth when sequencing the entire schistosome genome requires two samples per lane, costing around $1000 per sample (library preparation cost is, in this case, negligible). Capturing and sequencing exomes
<table>
<thead>
<tr>
<th>Library</th>
<th>Total reads</th>
<th>Mapped reads</th>
<th>% Mapped reads</th>
<th>Paired reads</th>
<th>Singletons</th>
<th>Duplicates</th>
<th>% Bait regions captured</th>
<th>Mean (median) read depth in bait regions</th>
<th>Mean (median) read depth in bait regions + adjacent 250 bp</th>
<th>ZW ratio</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sm.TZ_009.10.1</td>
<td>25 798 662</td>
<td>25 419 182</td>
<td>98</td>
<td>23 340 898</td>
<td>75 658</td>
<td>778 686</td>
<td>99.7</td>
<td>111.6 (50)</td>
<td>29.46 (11)</td>
<td>1.09</td>
<td>M</td>
</tr>
<tr>
<td>Sm.TZ_009.11.1</td>
<td>10 584 471</td>
<td>8 990 428</td>
<td>84</td>
<td>8 401 922</td>
<td>30 943</td>
<td>472 062</td>
<td>99.61</td>
<td>41.13 (30)</td>
<td>80.07 (17)</td>
<td>1.1</td>
<td>M</td>
</tr>
<tr>
<td>Sm.TZ_009.2.2</td>
<td>6 322 519</td>
<td>6 285 705</td>
<td>99</td>
<td>5 908 458</td>
<td>18 458</td>
<td>108 142</td>
<td>99.57</td>
<td>28.25 (22)</td>
<td>20.44 (9)</td>
<td>0.56</td>
<td>F</td>
</tr>
<tr>
<td>Sm.TZ_009.4.2</td>
<td>14 336 010</td>
<td>13 867 259</td>
<td>96</td>
<td>12 681 806</td>
<td>44 233</td>
<td>398 843</td>
<td>99.45</td>
<td>61.84 (29)</td>
<td>45.18 (11)</td>
<td>0.7</td>
<td>F</td>
</tr>
<tr>
<td>Sm.TZ_009.5.2</td>
<td>12 072 285</td>
<td>11 756 999</td>
<td>97</td>
<td>10 697 666</td>
<td>34 578</td>
<td>270 333</td>
<td>99.02</td>
<td>52.42 (24)</td>
<td>38.6 (9)</td>
<td>0.62</td>
<td>F</td>
</tr>
<tr>
<td>Sm.TZ_009.6.1</td>
<td>6 001 101</td>
<td>5 932 580</td>
<td>98</td>
<td>5 457 864</td>
<td>19 447</td>
<td>94 657</td>
<td>99.52</td>
<td>25.97 (20)</td>
<td>19.02 (8)</td>
<td>1.15</td>
<td>M</td>
</tr>
<tr>
<td>Sm.TZ_009.7.1</td>
<td>14 570 100</td>
<td>14 433 095</td>
<td>99</td>
<td>13 250 586</td>
<td>41 663</td>
<td>245 158</td>
<td>99.73</td>
<td>59.91 (45)</td>
<td>43 (14)</td>
<td>0.61</td>
<td>F</td>
</tr>
<tr>
<td>Sm.TZ_009.8.1</td>
<td>7 832 169</td>
<td>7 306 500</td>
<td>93</td>
<td>6 876 540</td>
<td>20 661</td>
<td>137 371</td>
<td>99.6</td>
<td>32.86 (25)</td>
<td>23.72 (10)</td>
<td>1.08</td>
<td>M</td>
</tr>
<tr>
<td>Sh.TZ_PEM0063.1</td>
<td>7 029 068</td>
<td>6 464 873</td>
<td>91</td>
<td>6 040 332</td>
<td>7279</td>
<td>179 216</td>
<td>99.73</td>
<td>28.72 (22)</td>
<td>20.55 (9)</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Sh.TZ_PEM0076.1</td>
<td>11 544 736</td>
<td>11 392 803</td>
<td>98</td>
<td>9 696 738</td>
<td>67 595</td>
<td>480 559</td>
<td>99.92</td>
<td>41.62 (35)</td>
<td>30.07 (12)</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Sh.TZ_PEM0079.1</td>
<td>5 557 924</td>
<td>4 998 091</td>
<td>89</td>
<td>4 672 272</td>
<td>6737</td>
<td>127 514</td>
<td>98.7</td>
<td>22.69 (14)</td>
<td>16.72 (6)</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Sh.TZ_PEM0089.2</td>
<td>18 364 494</td>
<td>17 933 751</td>
<td>97</td>
<td>16 683 002</td>
<td>20 437</td>
<td>390 640</td>
<td>99.85</td>
<td>71.88 (57)</td>
<td>51.08 (17)</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Sh.TZ_PEM0094.2</td>
<td>15 443 954</td>
<td>15 214 835</td>
<td>98</td>
<td>14 091 206</td>
<td>19 001</td>
<td>317 997</td>
<td>99.85</td>
<td>61.23 (52)</td>
<td>44.06 (17)</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Sh.TZ_PEM0099.2</td>
<td>7 069 391</td>
<td>6 672 191</td>
<td>94</td>
<td>6 292 150</td>
<td>7106</td>
<td>124 974</td>
<td>99.78</td>
<td>28.88 (24)</td>
<td>20.64 (9)</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Sh.TZ_PEM1013.1</td>
<td>7 958 910</td>
<td>7 809 025</td>
<td>98</td>
<td>7 293 706</td>
<td>28 039</td>
<td>487 198</td>
<td>99.8</td>
<td>31.72 (26)</td>
<td>22.86 (10)</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Sh.TZ_PEM1014.1</td>
<td>5 051 982</td>
<td>4 824 636</td>
<td>95</td>
<td>4 079 848</td>
<td>40 950</td>
<td>200 550</td>
<td>99.77</td>
<td>17.83 (15)</td>
<td>13.17 (6)</td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>

The ratio Z/W is the ratio of the read depth of the Z-linked regions (between 3.5–19.5 and 23.5–31 Mb) over the read depth of the rest of the Z chromosome. We considered a ratio between 0.5 and 0.75 corresponding to females, and a ratio over 0.75 to males. ND: not determined because the Schistosoma haematobium sexual chromosome is not identified in the current assembly.
also reduces both the analysis time and the storage capacity required for housing data because the exome represents just 4% of the complete genome (15 vs 365 Mb genome). Exome capture is currently about 4-fold less expensive than whole genome sequencing (RADseq) at present, ignoring the additional costs for data storage and analysis of whole genome data. The central advantage of exome capture over both RADseq and whole genome sequencing is that contaminating sequences can be effectively removed.

Our established workflow using the WGA method and qPCR quality check (Fig. 1) will facilitate the transition from population genetics using a handful of loci, to population genomics using genome-wide information. Using these tools, we will be able to fully exploit large collections of archived miracidia, such as those available through SCAN. For example, analysis of archived miracidia collected may be particularly useful for retrospective analyses of loci underlying drug resistance (Chevalier et al., 2016). This same approach can also be used for microscopic larval stages of other parasite species.

Table 3. Variant calling statistics

<table>
<thead>
<tr>
<th>Type of site</th>
<th>Number of sites in Schistosoma mansoni</th>
<th>Number of sites in Schistosoma haematobium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable sites</td>
<td>117 007</td>
<td>80 483</td>
</tr>
<tr>
<td>before filtering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variable sites</td>
<td>85 133</td>
<td>60 193</td>
</tr>
<tr>
<td>after filtering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biallelic sites</td>
<td>81 556</td>
<td>57 897</td>
</tr>
<tr>
<td>Multiallelic sites</td>
<td>443</td>
<td>522</td>
</tr>
<tr>
<td>Sites with insertion</td>
<td>1335</td>
<td>933</td>
</tr>
<tr>
<td>Sites with deletion</td>
<td>2038</td>
<td>1072</td>
</tr>
</tbody>
</table>

Number of variable sites identified in the eight exome capture libraries of S. mansoni and S. haematobium. We used conservative filtering parameters: minimum read depth of 10 and a minimum genotype quality (GQ) of 80.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0031182018000811

Acknowledgements. We gratefully acknowledge the help and support of our collaborator in Senegal, the late Oumar T. Diaw.

Financial support. Work at Texas Biomedical Research Institute (TBRI) was supported by the National Institute of Allergy and Infectious Diseases Grants (R01 AI097576-01 and SR21AI96277-01), by the National Center for Advancing Translational Sciences (CTSA/IIMS award no. UL1TR001120) and by Texas Biomedical Research Institute Forum grant (award 0467) and was conducted in facilities constructed with support from Research Facilities Improvement Program grant C06 RR013556 and RR017515 from the National Center for Research Resources of the National Institutes of Health. WL was supported by a Cowles Fellowship. CONTRAST was funded by the European Commission (FP6 STREP contract no: 032203). SCORE (https://score.uga.edu/) collecting activities in Tanzania and Zanzibar were funded by the University of Georgia Research Foundation Inc. (prime award no. 50816, sub awards RR374-053/4785416 and RR374-053/4893206), which is funded by the Bill & Melinda Gates Foundation for SCORE projects. SCAN is funded with support from the Wellcome Trust (grant no. 104958/Z/14/Z).

Conflict of interest. None.

References

Aemero M et al. (2015) Genetic diversity, multiplicity of infection and population structure of Schistosoma mansoni isolates from human hosts in Ethiopia. BMC Genetics 16, 137

https://doi.org/10.1017/S0031182018000811 Published online by Cambridge University Press


Ezeamama AE et al. (2016) Gaining and sustaining schistosomiasis control: study protocol and baseline data prior to different treatment strategies in five African countries. *MCM Infectious Diseases* 16, 229.


Gower CM et al. (2017) Phenotypic and genotypic monitoring of *Schistosoma mansoni* in Tanzanian schoolchildren five years into a preventative chemotherapy national control programme. *Parasites & Vectors* 10, 593.


Shaw ND et al. (2017) SMCHD1 mutations associated with a rare muscular dystrophy can also cause isolated arthria and Bosma arhinia microphthalmia syndrome. *Nature Genetics* 49, 238–248.


Valentim CLL et al. (2009) Efficient genotyping of *Schistosoma mansoni* miracidia following whole genome amplification. *Molecular and Biochemical Parasitology* 166, 81–84.


