Parasitology

cambridge.org/par

Whole genome amplification and exome sequencing of archived schistosome miracidia

Winka Le Clec'h[1], Frédéric D. Chevalier[1], Marina McDew-White[1], Fiona Allan[2], Bonnie L. Webster[2], Anouk N. Gouvras[2], Safari Kinunghi[3], Louis-Albert Tchuem Tchuenté[4,5], Amadou Garba[6], Khalfan A. Mohammed[7], Shaali M. Ame[8], Joanne P. Webster[9], David Rollinson[2], Aidan M. Emery[2] and Timothy J. C. Anderson[1]

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

Special Issue Research Article

Cite this article: Le Clec'h W *et al* (2018). Whole genome amplification and exome sequencing of archived schistosome miracidia. *Parasitology* **145**, 1739–1747. https://doi.org/10.1017/S0031182018000811

Received: 31 January 2018
Revised: 27 March 2018
Accepted: 5 April 2018
First published online: 28 May 2018

Key words:
Exome sequencing; FTA cards; miracidia; quantitative PCR; *Schistosoma*; whole genome amplification

Author for correspondence:
Winka Le Clec'h, E-mail: winkal@txbiomed.org

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 *Schistosome 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 cercaria 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

CAMBRIDGE UNIVERSITY PRESS

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/HQ/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).

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

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 Illustra 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 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 (v2.2.29) 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 AB1 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 C_T 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*: CGAAATTGGAG TTTGCTGTGT; *S. haematobium*: GGTGGTACTGGTTCTGG TTT) and 10 μM reverse primer (*S. mansoni*: TGTAGGTTGGAC GCTCTATATC; *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}: 2.19×10^1, 2.19×10^2, 2.19×10^3, 2.19×10^4, 2.19×10^5, 2.19×10^6, 2.19×10^7) or seven dilutions of a purified *α-tubulin 2* PCR product for *S. haematobium* (*α-tubulin 2* copies.μL^{-1}: 1.29×10^1, 1.29×10^2, 1.29×10^3, 1.29×10^4, 1.29×10^5, 1.29×10^6, 1.29×10^7). 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 SureSelect^{XT2} 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, Massachussetts, 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 SureSelect^{XT} Target Enrichment System (Agilent, Lexington, Massachussetts, 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 *ShSULT-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.

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

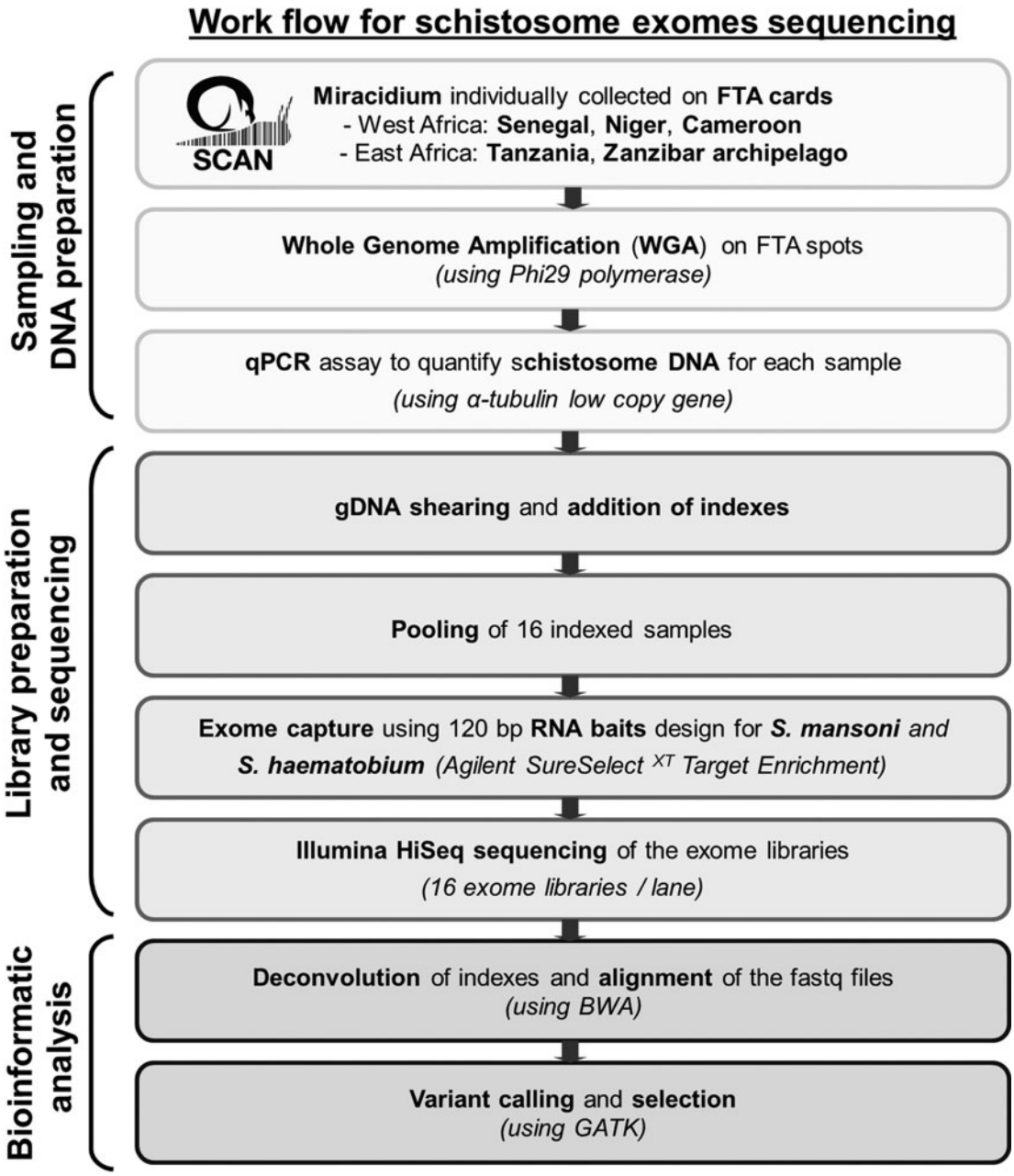

Fig. 1. Workflow summarizing the protocol used to generate and analyse genomic data starting with field collecting miracidium.

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 from East ($n = 60$) and West Africa ($n = 37$). Two samples from each region (3% of East African and 5% West African samples) contained <100 copies. WGA material from East African samples showed higher quantities of *S. mansoni* DNA than WGA material from West African samples (Fig. 2; Wilcoxon test; $P = 1.463 \times 10^{-6}$).

We observed similar heterogeneity in schistosome DNA quantity for WGA from the 65 *S. haematobium*-positive samples from

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

Table 1. Whole genome amplification and qPCR assay summary statistics

Schistosome spp.	Location	Country	Collection date	No. schistosome positive WGA/total no. WGA performed	% Schistosome-positive samples by qPCR	No. of patients with schistosome-positive WGA/no. of patients	Average no. of spots tested per patient[a]	Average (+S.D.) DNA conc. after WGA (ng μL^{-1})[b]
S. mansoni	E. Africa	Tanzania	2012	60/101	59.40%	60/64	1.68	34 ± 5.19
	W. Africa	Senegal	2007–8	24/48	50%	23/27	2.09	28.74 ± 5.08
		Niger	2006–7	13/28	46.42%	13/17	2.23	24.63 ± 5.04
		Cameroon	2007	0/14	0%	0/6	–	–
S. haematobium	E. Africa	Zanzibar	2011	65/138	47.10%	62/80	2.22	22.45 ± 4.65
	W. Africa	Niger	2013	55/83	66.26%	52/57	1.60	21.21 ± 2.85

Summary of the number and percentage of positive schistosome samples, the number of patients from which schistosome-positive WGA were obtained, the average number of spots tested per patients in order to get at least one positive for schistosome, and the average (+S.D.) DNA concentration recovered after WGA in ng μL^{-1} in 130 μL of sample.

[a]We initially conducted WGA from a single miracidia per patient: further WGA were done only if the initial sample was schistosome-negative by qPCR.

[b]Only shown for schistosome-positive samples.

East Africa, while the quantity of DNA from 55 West African *S. haematobium* WGAs varied by only three orders of magnitude, with no outliers. Only 11% (7/65) of *S. haematobium*-positive samples, all from East Africa, showed <100 copies. We found a marginally higher quantity of *S. haematobium* DNA in West compared with East African samples (Fig. 2; Wilcoxon test; $P = 0.017$).

WGA from *S. haematobium* samples contained more schistosome DNA than from *S. mansoni* samples in both East (Fig. 2; Wilcoxon test; $P = 1.980 \times 10^{-7}$) and West African samples (Fig. 2; Wilcoxon test; $P = 2.472 \times 10^{-15}$).

Exome sequencing of single miracidia

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 (95.5 and 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 *ZW*). 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 revealed more variable sites in *S. mansoni* than in *S. haematobium*. The raw data were further filtered using very stringent parameters (minimum read depth of 10 and genotype quality of 80). After filtering, we retained informative sites accounting for around 74% of the initial sites identified. Among them, the vast majority (96%)

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

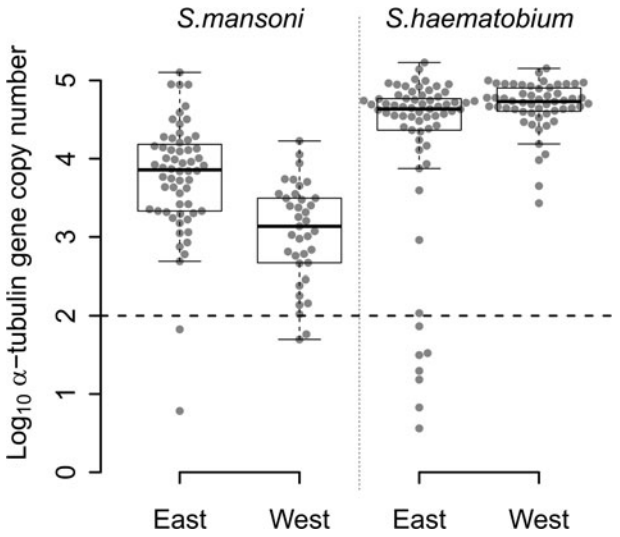

Fig. 2. Distribution of the number of *α-tubulin* copy among all the WGA schistosome samples tested. Log_{10} distribution of the number of *α-tubulin* copy in 20 ng of total DNA, among the 60 East African (Tanzania) and the 37 West African (Senegal and Niger) *Schistosoma mansoni* samples, and the 65 East African (Zanzibar) and 55 West African (Niger) *Schistosoma haematobium* samples from FTA-preserved samples. *Schistosome mansoni* and *S. haematobium* DNA varies in concentration by three to five orders of magnitude within WGA products. The 11 outliers, below the dotted line threshold, exhibited a very low number of *α-tubulin* copies (<100 copies in 20 ng of total DNA) and were removed from the next generation sequencing sample set.

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 WGAs, 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, minimalizing 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 (~US$900 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

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

Table 2. Exome capture library statistics

Library	Total reads	Mapped reads	% Mapped reads	Paired reads	Singletons	Duplicates	% Bait regions captured	Mean (median) read depth in bait regions	Mean (median) read depth in bait regions + adjacent 250 bp	*ZW* ratio	Sex
Sm.TZ_009.10.1	25 798 662	25 419 182	98	23 340 898	75 658	778 686	99.7	111.6 (50)	29.46 (11)	1.09	M
Sm.TZ_009.1.1	10 584 471	8 990 428	84	8 401 922	30 943	472 062	99.61	41.13 (30)	80.07 (17)	1.1	M
Sm.TZ_009.2.2	6 322 519	6 285 705	99	5 908 458	18 458	108 142	99.57	28.25 (22)	20.44 (9)	0.56	F
Sm.TZ_009.4.2	14 336 010	13 867 259	96	12 681 806	44 233	398 843	99.45	61.84 (29)	45.18 (11)	0.7	F
Sm.TZ_009.5.2	12 072 285	11 756 999	97	10 697 666	34 578	270 333	99.02	52.42 (24)	38.6 (9)	0.62	F
Sm.TZ_009.6.1	6 001 101	5 932 580	98	5 457 864	19 447	94 657	99.52	25.97 (20)	19.02 (8)	1.15	M
Sm.TZ_009.7.1	14 570 100	14 433 095	99	13 250 586	41 663	245 158	99.73	59.91 (45)	43 (14)	0.61	F
Sm.TZ_009.8.2	7 832 169	7 306 500	93	6 876 540	20 661	137 371	99.6	32.86 (25)	23.72 (10)	1.08	M
Sh.TZ_PEM0063.1	7 029 068	6 464 873	91	6 040 332	7279	179 216	99.73	28.72 (22)	20.55 (9)	ND	ND
Sh.TZ_PEM0076.1	11 544 736	11 392 803	98	9 696 738	67 595	480 559	99.92	41.62 (35)	30.07 (12)	ND	ND
Sh.TZ_PEM0079.1	5 557 924	4 998 091	89	4 672 272	6737	127 514	98.7	22.69 (14)	16.72 (6)	ND	ND
Sh.TZ_PEM0089.2	18 364 494	17 933 751	97	16 683 002	20 437	390 640	99.85	71.88 (57)	51.08 (17)	ND	ND
Sh.TZ_PEM0094.2	15 443 954	15 214 835	98	14 091 206	19 001	317 997	99.85	61.23 (52)	44.06 (17)	ND	ND
Sh.TZ_PEM0099.2	7 069 391	6 672 191	94	6 292 150	7106	124 974	99.78	28.88 (24)	20.64 (9)	ND	ND
Sh.TZ_PEM0103.1	7 958 910	7 809 025	98	7 293 706	28 039	487 198	99.8	31.72 (26)	22.86 (10)	ND	ND
Sh.TZ_PEM0104.1	5 051 982	4 824 636	95	4 079 848	40 950	200 550	99.77	17.83 (15)	13.17 (6)	ND	ND

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.

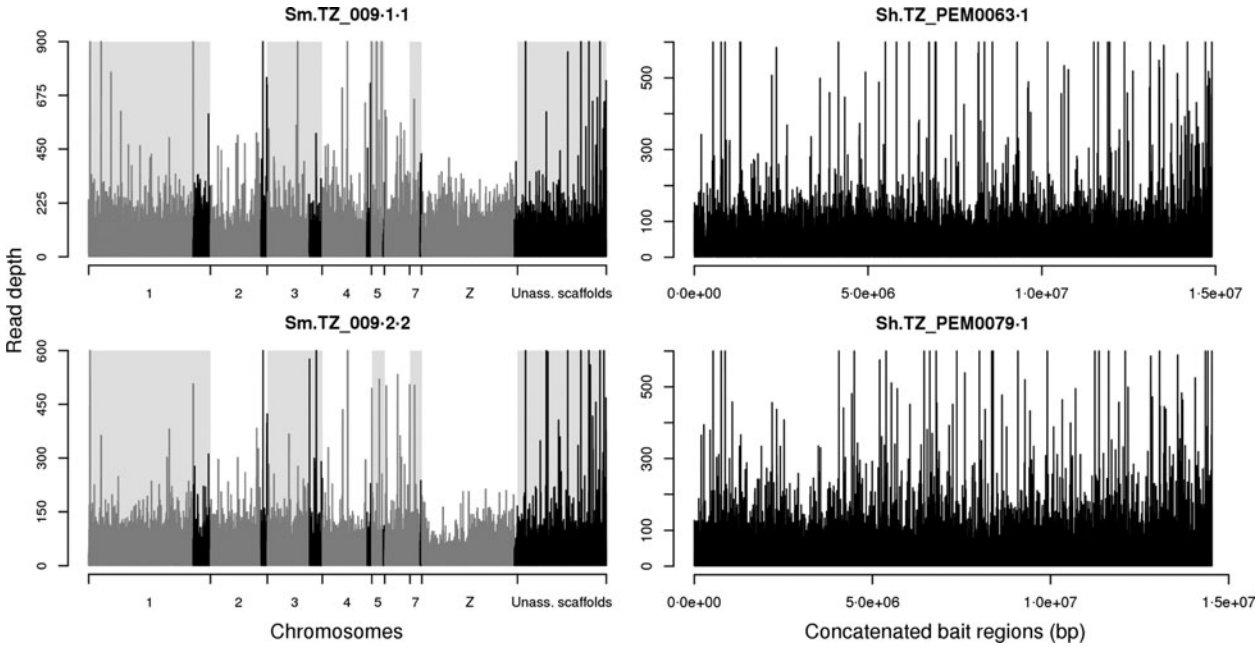

Fig. 3. Read depth of the bait regions for *Schistosoma mansoni* and *Schistosoma haematobium* single miracidium libraries. *Schistosoma mansoni* (Sm) plots show read depth of bait regions on the assembled chromosomes (grey), on unplaced scaffolds that have been assigned to chromosomes or unassigned scaffolds (black). *Schistosoma haematobium* (Sh) plots show read depth of the concatenated bait regions (bp, base pair).

Table 3. Variant calling statistics

Type of site	Number of sites in *Schistosoma mansoni*	Number of sites in *Schistosoma haematobium*
Variable sites before filtering	117 007	80 483
Variable sites after filtering	85 133	60 193
Biallelic sites	81 556	57 897
Multiallelic sites	443	522
Sites with insertion	1335	933
Sites with deletion	2038	1072

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.

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 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.

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 5R21AI096277-01), 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

Attard CRM, Beheregaray LB and Möller LM (2017) Genotyping-by-sequencing for estimating relatedness in non-model organisms: avoiding the trap of precise bias. *Molecular Ecology Resources* **12**, 381–390.

Berriman M *et al.* (2009) The genome of the blood fluke *Schistosoma mansoni*. *Nature* **460**, 352–358.

Biesecker LG, Shianna KV and Mullikin JC (2011) Exome sequencing: the expert view. *Genome Biology* **12**, 128.

Chevalier FD *et al.* (2014) Efficient linkage mapping using exome capture and extreme QTL in schistosome parasites. *BMC Genomics* **15**, 617.

Chevalier FD *et al.* (2016) Independent origins of loss-of-function mutations conferring oxamniquine resistance in a Brazilian schistosome population. *International Journal for Parasitology* **46**, 417–424.

Danecek P *et al.* (2011) The variant call format and VCF tools. *Bioinformatics (Oxford, England)* **27**, 2156–2158.

DePristo MA *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–498.

Duvaux-Miret O *et al.* (1991) Molecular cloning and sequencing of the α-tubulin gene from *Schistosoma mansoni*. *Molecular and Biochemical Parasitology* **49**, 337–340.

Emery AM *et al.* (2012) Schistosomiasis collection at NHM (SCAN). *Parasites & Vectors* **5**, 185.

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

Glenn TC *et al.* (2013) Significant variance in genetic diversity among populations of *Schistosoma haematobium* detected using microsatellite DNA loci from a genome-wide database. *Parasites & Vectors* **6**, 300.

Gower CM *et al.* (2007) Development and application of an ethically and epidemiologically advantageous assay for the multi-locus microsatellite analysis of *Schistosoma mansoni*. *Parasitology* **134**, 523.

Gower CM *et al.* (2011) Population genetics of *Schistosoma haematobium*: development of novel microsatellite markers and their application to schistosomiasis control in Mali. *Parasitology* **138**, 978–994.

Gower CM *et al.* (2013) Population genetic structure of *Schistosoma mansoni* and *Schistosoma haematobium* from across six sub-Saharan African countries: implications for epidemiology, evolution and control. *Acta Tropica* **128**, 261–274.

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.

Harvey MG *et al.* (2016) Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Systematic Biology* **65**, 910–924.

Hotez PJ *et al.* (2006) Helminth infections: soil-transmitted helminth infections and Schistosomiasis. In Jamison DT, Breman JG, Measham AR, Alleyne G, Claeson M, Evans DB, Jha P, Mills A and Musgrove P (eds), *Disease Control Priorities in Developing Countries*. Washington, DC: World Bank, Chapter 24.

Lepesant JM *et al.* (2012) Chromatin structure changes around satellite repeats on the *Schistosoma mansoni* female sex chromosome suggest a possible mechanism for sex chromosome emergence. *Genome Biology* **13**, R14.

Li H and Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–1760.

Li H *et al.* (2009) The sequence alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078–2079.

Li Y *et al.* (2017) Genetic diversity and selection of three nuclear genes in *Schistosoma japonicum* populations. *Parasites & Vectors* **10**, 87.

McKenna A *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303.

Pehlivan D *et al.* (2014) Whole-exome sequencing links TMCO1 defect syndrome with cerebro-facio-thoracic dysplasia. *European Journal of Human Genetics* **22**, 1145–1148.

Protasio AV *et al.* (2012) A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS Neglected Tropical Diseases* **6**, e1455.

Quinlan AR and Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841–842.

Rabbani B, Tekin M and Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics* **59**, 5–15.

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

Shortt JA *et al.* (2017) Whole genome amplification and reduced-representation genome sequencing of *Schistosoma japonicum* miracidia. *PLoS Neglected Tropical Diseases* **11**, e0005292.

Steinauer ML, Blouin MS and Criscione CD (2010) Applying evolutionary genetics to schistosome epidemiology. *Infection, Genetics and Evolution* **10**, 433–443.

Steinauer ML *et al.* (2013) Non-invasive sampling of schistosomes from humans requires correcting for family structure. *PLoS Neglected Tropical Diseases* **7**, e2456.

Taylor AB *et al.* (2017) Structural and enzymatic insights into species-specific resistance to schistosome parasite drug therapy. *Journal of Biological Chemistry* **292**, 11154–11164.

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

Van den Broeck F *et al.* (2015) Reconstructing colonization dynamics of the human parasite *Schistosoma mansoni* following anthropogenic environmental changes in Northwest Senegal. *PLoS Neglected Tropical Diseases* **9**, e0003998.

Visser PS and Pitchford RJ (1972) A simple apparatus for rapid recovery of helminth eggs from excreta with special reference to *Schistosoma mansoni*. *South African Medical Journal=Suid-Afrikaanse tydskrif vir geneeskunde* **46**, 1344–1346.

Webster PJ *et al.* (1992) A cDNA encoding an α-tubulin from *Schistosoma mansoni*. *Molecular and Biochemical Parasitology* **51**, 169–170.

Webster BL *et al.* (2012) Genetic diversity within *Schistosoma haematobium*: DNA barcoding reveals two distinct groups. *PLoS Neglected Tropical Diseases* **6**, e1882.

Webster BL *et al.* (2013) DNA 'barcoding' of *Schistosoma mansoni* across sub-Saharan Africa supports substantial within locality diversity and geographical separation of genotypes. *Acta Tropica* **128**, 250–260.

Webster BL *et al.* (2015) Development of novel multiplex microsatellite polymerase chain reactions to enable high-throughput population genetic studies of *Schistosoma haematobium*. *Parasites & Vectors* **8**, 432.

Young ND *et al.* (2012) Whole-genome sequence of *Schistosoma haematobium*. *Nature Genetics* **44**, 221–225.

Zhou Y *et al.* (2009) The *Schistosoma japonicum* genome reveals features of host–parasite interplay. *Nature* **460**, 345–351.

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