Evaluation of DNA extraction methods on individual helminth egg and larval stages for whole genome sequencing

Stephen R. Doyle[1*], Geetha Sankaranarayan[1], Fiona Allen[2], Duncan Berger[1], Pablo D. Jimenez Castro[3], James Bryant Collins[3], Thomas Crellen[1,10], María A. Duque-Correa[1], Peter Ellis[1], Tegegn G. Jaleta[3], Roz Laing[4], Kirsty Maitland[4], Catherine McCarthy[1], Tchonfienet Moundai[6], Ben Softley[1], Elizabeth Thiele[7], Philippe Tchindebet Ouakou[6], John Vianney Tushabe[1,11], Joanne P. Webster[8], Adam J. Weiss[9], James Lok[4], Eileen Devaney[5], Ray M. Kaplan[3], James A. Cotton[1], Matthew Berriman[1], Nancy Holroyd[1*]

* Correspondence:
Stephen R. Doyle: stephen.doyle@sanger.ac.uk
Nancy Holroyd: neh@sanger.ac.uk

[1.] Wellcome Sanger Institute, Hinxton, Cambridgeshire, United Kingdom
[2.] Department of Life Sciences, Natural History Museum, London, United Kingdom
[3.] Department of Infectious Diseases, College of Veterinary Medicine, University of Georgia, Athens, Georgia, USA
[4.] Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[5.] Institute of Biodiversity Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Garscube Campus, Glasgow, United Kingdom
[6.] Department of Biology, Vassar College, Poughkeepsie, New York, USA
[7.] Ministry of Public Health, N'Djamena, Chad
[8.] Centre for Emerging, Endemic and Exotic Diseases, Department of Pathology and Population Sciences, Royal Veterinary College, University of London, Herts, United Kingdom
[9.] Guinea Worm Eradication Program, The Carter Center, Atlanta, Georgia, USA
[10.] Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[11.] Medical Research Council/Uganda Virus Research Institute and London School of Hygiene & Tropical Medicine Uganda Research Unit, P.O. Box 49, Entebbe, Uganda

Keywords: Helminth, genomics, whole genome sequencing, DNA extraction, low input, diagnostics

Abstract

Whole genome sequencing is being rapidly applied to the study of helminth genomes, including *de novo* genome assembly, population genetics, and diagnostic

applications. Although late-stage juvenile and adult parasites typically produce sufficient DNA for molecular analyses, these parasitic stages are almost always inaccessible in the live host; immature life stages found in the environment for which samples can be collected non-invasively offer a potential alternative, however, these samples are typically yield very low quantities of DNA, can be environmentally resistant, and are susceptible to contamination, often from bacterial or host DNA. Here, we have tested five low-input DNA extraction protocols together with a low-input sequencing library protocol to assess the feasibility of whole genome sequencing of individual immature helminth samples. These approaches do not use whole genome amplification, a common but costly approach to increase the yield of low input samples. We first tested individual parasites from two species spotted onto FTA cards - egg and L1 stages of *Haemonchus contortus* and miracidia of *Schistosoma mansoni* - before further testing on an additional six species - *Ancylostoma caninum, Ascaridia dissimilis, Dirofilaria immitis, Dracunculus medinensis, Strongyloides stercoralis, and Trichuris muris* - with an optimal protocol. Whole genome sequencing followed by analyses to determine the proportion of on- and off-target mapping revealed successful sample preparations for six of the eight species tested with variation between species, and within species but between life stages, described. These results demonstrate the feasibility of whole genome sequencing of individual parasites, and highlight a new avenue towards generating sensitive, specific, and information-rich data for the diagnosis and surveillance of helminths.

68 **Introduction**

69 Accurate methods for diagnosis and surveillance of helminth infections are of
70 increasing interest in both human and animal health settings. Such approaches are
71 typically proposed to monitor the presence and ultimately decline of populations
72 targeted by large-scale control measures, such as mass drug administration (MDA)
73 for the prevention and/or treatment of human helminth infections, or prophylactic
74 treatment of domesticated animals. An ideal diagnostic will be sensitive, to detect
75 the parasite if in fact present, and specific, to identify the targeted parasite species
76 in the presence of non-target material such as other parasite species or the host.
77 Ideally, samples taken for diagnostic purposes can be used to gather additional
78 information beyond the presence or absence of a specific parasite, so the same
79 material could be used for example, to predict how well the infection will respond to
80 drug treatment, or how the parasite is related to other endemic or imported
81 parasites. As most parasitic stages of helminths of humans and animals are naturally
82 inaccessible *in vivo* (not accounting for potential availability of some mature stages
83 of helminths following chemo-expulsion, for example, *Ascaris lumbricoides* and
84 *Trichuris trichiura*), a diagnostic should also be informative on non-invasive stages of
85 the parasite, such as eggs deposited in faeces, or intermediate stages of the
86 parasite's life cycle that exist in the external environment.

87

88 A key challenge of working with environmental intermediate stages is that they are
89 often immature, for example, eggs or early stage larvae, and extremely small (for
90 example, *Haemonchus contortus* eggs are approximately 75 × 44 µm and
91 *Schistosoma mansoni* miracidia approximately 140 × 55 µm), limiting the amount of
92 accessible material (e.g., DNA) available to perform the diagnostic assay. They are
93 often environmentally resistant, and the same features that naturally protect the
94 DNA from damage prior to reinfection make it difficult to extract DNA. In many cases
95 they are isolated from host faeces and so are susceptible to bacterial contamination,
96 or from host tissues and so become contaminated with host DNA. Furthermore,
97 samples may need to be transported efficiently to a laboratory setting without a
98 significant loss of this already limited material. A number of approaches have been
99 tested to preserve macromolecules from individual parasites for transport and

100 storage, including ethanol, RNAlater and Whatman® FTA® cards, from which robust
101 PCR and microsatellite data could be profiled (Boué et al., 2017; Campbell et al.,
102 2017; Gower et al., 2007; Marek et al., 2014; Webster, 2009; Webster et al., 2012;
103 Xiao et al., 2013). Although under ideal conditions the detection of a single DNA
104 molecule is possible, the limited material available per parasite has to date largely
105 restricted assaying to a small number of loci, limiting the amount of information
106 obtained from any individual parasite.
107
108 Genomic approaches offer an information-rich technology for diagnostic and
109 surveillance applications. Increasing throughput and decreasing costs of whole
110 genome sequencing has resulted in the recent and steadily growing application of
111 genomics in helminth parasitology; for example, for diagnostic applications, high
112 throughput amplicon sequencing for helminth species identification and community
113 composition (Avramenko et al., 2015) and the presence of drug resistance alleles
114 (Avramenko et al., 2019) have been described. Although low DNA concentrations are
115 typically prohibitive for genome-wide approaches on individual parasites, a number
116 of studies have successfully used whole genome amplification on DNA extracted
117 from single larval stages to perform reduced representation (Shortt et al., 2017) and
118 exome (Le Clec'h et al., 2018; Platt et al., 2019) sequencing on miracidia of
119 *Schistosoma* spp., and whole genome sequencing of *Haemonchus contortus* L3 stage
120 larvae (Doyle et al., 2018) and microfilaria of *Wuchereria bancrofti* (Small et al.,
121 2018). Whole genome amplification protocols do, however, add considerable
122 expense per sample, and can introduce technical artefacts such as uneven and/or
123 preferential amplification (potentially of contaminant sequences), chimeric
124 sequences, and allele dropout (Sabina and Leamon, 2015; Tsai et al., 2014), that may
125 lead to a reduction in genetic diversity, and in turn, relevance to the original
126 unamplified material. The field of genomics is, however, rapidly advancing towards
127 very low minimum sample input requirements, and single cell approaches for DNA
128 and RNA sequencing are now available. Such approaches have begun to be used on
129 parasitic species such as *Plasmodium* spp. (Howick et al., 2019; Ngara et al., 2018;
130 Reid et al., 2018; Trevino et al., 2017), but are yet to be adopted by helminth
131 parasitologists. Although these low-input, high-throughput approaches are not

132 designed – and perhaps not currently suitable – for diagnostic applications, the
133 developments in molecular biology techniques for low input sequencing can benefit
134 the use of genomics for helminth applications. Here, we test a number of low-input
135 DNA extraction approaches for individual helminth samples stored on Whatman®
136 FTA® cards, followed by low-input library preparation without whole genome
137 amplification, and whole-genome sequencing. A total of five DNA extraction
138 approaches were initially tested, after which the most promising approach was
139 applied to multiple life stages from eight helminth species. The results presented
140 here demonstrate the advancement of low-input whole genome sequencing, and are
141 discussed in the context of their utility for helminth diagnostics and surveillance.
142
143

144 **Methods**

145 **Sample collection**

146 Samples representing accessible, immature life stages of a total of eight helminth
147 species were tested, the collection of which is described below.

148

149 *Ancylostoma caninum:* Fresh feces from a research purpose-bred laboratory beagle
150 (University of Georgia AUP # A2017 10-016-Y1-A0) infected with the Barrow isolate
151 (drug-susceptible isolate from Barrow County Georgia, USA) were collected and
152 made into a slurry with water, filtered through 425 µm and 180 µm sieves, and
153 centrifuged at 2500 rpm for 5 min after which the supernatant was discarded. Kaolin
154 (Sigma-Aldrich, St. Louis, MO) was then added and resuspended in sodium nitrate
155 (SPG 1.25–1.3) (Feca-Med®, Vedco, Inc. St Joseph, MO, USA). The tube was then
156 centrifuged at 2500 rpm for 5 min, after which the supernatant was passed through
157 a 30 µm sieve and rinsed with distilled water, and reduced to a volume of 10-15 mL.
158 The volume was adjusted to 1 egg per 5 µL using distilled water. The eggs were
159 stored at room temperature for 2 h before placing them onto the Whatman® FTA®
160 cards. Eggs were also placed onto Nematode growth medium (NGM) plates (Sulston
161 and Hodgkin, 1988) and incubated at 26°C to obtain the first-stage (L1) larvae. After
162 48 h, larvae were rinsed off the plate with distilled water and centrifuged at 1000
163 rpm for 5 mins. Larvae were counted and the concentration adjusted to 1 larva per 5
164 µL. The larvae were stored at room temperature for 2 h before placing them onto
165 the Whatman® FTA® cards. To obtain third-stage (L3) larvae, eggs were isolated from
166 fresh feces from a research purpose-bred laboratory beagle (University of Georgia
167 AUP # A2017 10-016-Y1-A0) infected with the Worthy isolate (Worthy 3.1F3Pyr;
168 multiple-drug resistant isolate originally isolated from a greyhound dog, Florida,
169 USA). Eggs were placed onto NGM plates (Sulston and Hodgkin, 1988) and incubated
170 at 26°C. After seven days, larvae were rinsed off the plate with distilled water and
171 centrifuged at 1000 rpm for 5 mins. Larvae were counted and the concentration
172 adjusted to 1 larva per 5 µL. The larvae were stored at room temperature for 2 h
173 before placing them onto the Whatman® FTA® cards.

174

175 *Ascaridia dissimilis*: Eggs of *A. dissimilis* (Isolate Wi: North Carolina, USA) were
176 isolated from excreta of experimentally infected turkeys. Water was added to the
177 excreta and made into a slurry, which was filtered using a 425 µm and 180 µm sieve
178 to remove large debris. The remaining particulates were placed into 50 mL
179 centrifuge tubes and centrifuged at 433 g for 7 mins. Supernatant was removed and
180 the pellet was resuspended in a saturated sucrose solution with a specific gravity of
181 1.15. The suspension was centrifuged as before and eggs were isolated from the top
182 layer. Eggs were rinsed over a 20 µm sieve with water to remove residual sucrose
183 and then concentrated to 1 egg per 5 µL using deionized water. Multiple 5 µL
184 aliquots of the egg solution were dispensed using a micropipette onto the
185 Whatman® FTA® card.
186
187 *Dirofilaria immitis*: Blood was taken from a dog infected with the macrocyclic lactone
188 (ML)-resistant Yazoo strain (Yazoo: originally isolated from a dog in Yazoo City,
189 Mississippi, USA; See Maclean *et al.* (2017) for complete history). To obtain
190 microfilariae, blood was collected in heparin tubes and centrifuged for 30 min at
191 2500 rpm after which the supernatant was discarded. The pellet was suspended in
192 3.8% sodium citrate (Sigma-Aldrich, St. Louis, MO) and 15% saponin (Sigma-Aldrich,
193 St. Louis, MO) was added in a 1:7 dilution. The tube was then vortexed and
194 centrifuged for 30 min at 2500 rpm after which the supernatant was discarded and
195 the pellet resuspended in 3.8% sodium citrate to the original blood volume,
196 vortexed, and then centrifuged for 4 min at 2500 rpm. The pellet was then
197 resuspended and mixed in a 1:9 solution of 10× phosphate buffered saline (PBS;
198 Thermo Fisher Scientific, Waltham, MA) and distilled water. The tube was then
199 centrifuged for 4 min at 2500 rpm and the pellet resuspended in PBS. The
200 microfilariae were then counted and adjusted accordingly to have one microfilaria
201 per 5 µL and stored at room temperature for 2 h before placing them onto the
202 Whatman® FTA® cards.
203
204 *Dracunculus medinensis*: Individual L1 samples were obtained as progeny of an adult
205 female worm manually extracted from an infected dog in Tarangara village, Chad
206 (9.068611 N, 18.708611 E) in 2016. This extraction forms part of the standard

207 containment and treatment procedure for Guinea worm infections, as agreed upon
208 and sanctioned by the World Health Organisation and country ministries of health.
209 The adult worm was submerged in ethanol in a microcentrifuge tube for storage; L1
210 stage progeny that were found settled on the bottom of the tube were collected for
211 analysis.
212
213 *Haemonchus contortus:* Eggs representing the F5 generation of a genetic cross
214 (described in Doyle *et al.* (2018)) were collected from fresh faeces from
215 experimentally infected sheep housed at the Moredun Research Institute, UK. All
216 experimental procedures were examined and approved by the Moredun Research
217 Institute Experiments and Ethics Committee and were conducted under approved
218 UK Home Office licenses (PPL 60/03899) in accordance with the Animals (Scientific
219 Procedures) Act of 1986. Briefly, faeces were mixed with tap water and passed
220 through a 210 µm sieve, then centrifuged at 2500 rpm for 5 mins in polyallomer
221 tubes. The supernatant was discarded, before adding kaolin to the faecal pellet,
222 vortexing, and resuspending in a saturated salt solution. After centrifugation at 1000
223 rpm for 10 mins, the polyallomer tube was clamped to isolate eggs, which were
224 collected on a 38 µm sieve and rinsed thoroughly with tap water. Eggs were
225 incubated on NGM plates at 20°C for 48 h to hatch to L1 stage larvae. In addition to
226 freshly collected material, eggs collected in the same manner then stored at -20°C,
227 from a previous generation of the cross, were also tested. Eggs and L1 larvae were
228 resuspended in PBS and spotted onto Whatman FTA cards in 3 µL per egg/L1.
229
230 *Schistosoma mansoni*: Three collections of *S. mansoni* samples were used in this
231 work. The first were field samples collected from humans on Lake Victoria fishing
232 villages in Uganda as part of the LaVIISWA trial (Sanya et al., 2018). Ethical approval
233 for this trail was given by the Uganda Virus Research Institute (reference number
234 GC127), Uganda National Council for Science and Technology (reference number HS
235 1183) and London School of Hygiene & Tropical Medicine (reference number 6187).
236 Parasite eggs were collected from participants' stool samples using a Pitchford-Visser
237 funnel, washed with mineral water until clean, and transferred into a petri dish with
238 water to be hatched in direct sunlight. After hatching, the miracidia were picked in 2

239 µL water using a pipette and placed on a Whatman® FTA® card for storage. The
240 second were field samples collected as part of a repeated cross-sectional study of
241 MDA exposure in school children in Uganda. Patient enrolment, including written
242 consent, and sample collection have been described previously (Crellen et al., 2016).
243 Ethical approvals for this study were granted by the Uganda National Council of
244 Science and Technology (MoU sections 1.4, 1.5, 1.6) and the Imperial College
245 Research Ethics Committee (EC NO: 03.36. R&D No: 03/SB/033E). Host stool was
246 sampled 1-3 days prior to treatment with praziquantel (40 mg/kg) and albendazole
247 (400 mg). A Pitchford-Visser funnel was used to wash and filter stool to retain
248 parasite eggs. The filtrate was kept overnight in water and hatched the following
249 morning in sunlight. Individual miracidia were isolated with a 20 µL pipette and
250 transferred into petri-dishes of nuclease-free water twice before spotting onto
251 Whatman® FTA® cards. The third source of miracidia were derived from the livers of
252 experimentally infected mice kept to maintain the *S. mansoni* life cycle at Wellcome
253 Sanger Institute. Mouse infection protocols were approved by the Welfare and
254 Ethical Review Body (AWERB) of the Wellcome Sanger Institute. The AWERB is
255 constituted as required by the UK Animals (Scientific Procedures) Act 1986
256 Amendment Regulations 2012. BalbC mice (6-8 week old) were infected with 250
257 cercariae, after which livers were collected on day 40 post infection. Eggs were
258 isolated from the liver tissues using collagenase digestion followed by percoll
259 gradient, and were washed well with sterile PBS, before being hatched in sterile
260 conditioned water. The hatched individual miracidia were spotted onto Whatman®
261 FTA® cards.
262
263 *Strongyloides stercoralis:* The *S. stercoralis* UPD strain and the isofemale isolate
264 PVO1 were maintained in purpose-bred, prednisone-treated mix breed dogs
265 according to protocol 804883 approved by the University of Pennsylvania
266 Institutional Animal Care and Use Committee (IACUC), USA. IACUC-approved
267 research protocols and all routine husbandry care of the animals was conducted in
268 strict accordance with the Guide for the Care and Use of Laboratory Animals of the
269 National Institutes of Health, USA. Faeces were collected, moisturized and mixed
270 with equal volume of charcoal and cultured at 22°C in 10 cm plates (Lok, 2007). Post

271 parasitic stage one and two larvae (L1/L2), free-living male and female adults, and
272 infective third-stage (L3) were isolated by the Baermann technique after 24 h, 48 h,
273 and 6 days in these charcoal coprocultures, respectively (Jaleta et al., 2017; Lok,
274 2007). For this study, free living females, L1 and L3 larvae were collected from the
275 Baermann funnel sediments and washed three times using PBS.
276
277 *Trichuris muris*: Infection and maintenance of *T. muris* was conducted as described
278 (Wakelin, 1967). The care and use of mice were in accordance with the UK Home
279 Office regulations (UK Animals Scientific Procedures Act 1986) under the Project
280 license P77E8A062 and were approved by the institutional Animal Welfare and
281 Ethical Review Body. Female SCID mice (6–10 wk old) were orally infected under
282 anaesthesia with isoflurane with a high dose (n = 400) of embryonated eggs from *T.*
283 *muris* E-isolate. Mice were monitored daily for general condition and weight loss. At
284 day 35 post infection, mice were killed by exsanguination under terminal anesthesia,
285 after which adult worms were harvested from cecums. Adult worms were cultured in
286 RPMI 1640 supplemented with 10% fetal calf serum (v/v), 2 mM L-glutamine,
287 penicillin (100 U/mL), and streptomycin (100 mg/mL; all Invitrogen), for 4 h or
288 overnight, and eggs were collected. The eggs were allowed to embryonate for at
289 least 6 weeks in distilled water, and infectivity was established by worm burden in
290 SCID mice. *T. muris* eggs were hatched to produce sterile L1 larvae using 32% sodium
291 hypochlorite in sterile water for 2 h at 37°C with 5% CO_2. Eggs were washed with
292 RPMI 1640 supplemented with 10% fetal calf serum (v/v), 2 mM L-glutamine,
293 penicillin (100 U/mL), and streptomycin (100 mg/mL; all Invitrogen), and incubated
294 at 37°C with 5% CO_2 for 4 to 5 days until they hatched.
295
296 For each species, unless otherwise stated, pools of individuals were washed in sterile
297 PBS, before being transferred to a petri dish. Individuals were identified under the
298 microscope, after which 5 µL PBS containing an individual parasite was transferred
299 onto a Whatman® FTA® card and dried for a minimum of 20 mins at room
300 temperature prior to storage or shipping to the Wellcome Sanger Institute, UK. The
301 Whatman® FTA® cards with samples spotted were stored in a clean plastic bag in the
302 dark at room temperature prior to analysis.

303

304

305 **DNA extraction**

306 The sample spots on FTA cards were punched out manually into 96-well plates using
307 either a Harris Punch or autonomously using robotics. The DNA extraction was
308 carried out for each method as described below:

309 1. *Nexttec* (NXT): Extraction using the Nexttec 1-step DNA Isolation Kit for
310 Tissues & Cells (cat: 10N.904; Waendel Technology Limited, UK) was
311 performed according to manufacturer's guidelines. 75 µl of proteinase lysis
312 buffer was used for digestion.
313 2. *Bloodspot* (BSP): The bloodspot extraction was derived from the QIAamp
314 DNA Investigator Kit (cat: 51104/51106; Qiagen), following the "DNA
315 Purification from Dried Blood Spots (QIAamp DNA Mini Kit)" protocol
316 according to manufacturer's guidelines.
317 3. *CGP* (CGP): Extraction using the CGP protocol (Moore et al., 2018) involved
318 adding 30 µL of lysis buffer (1.25 µg/mL of Protease reagent (Qiagen; cat#
319 19155) in Tris HCl pH 8.0, 0.5% Tween 20, 0.5% NP40) to individual
320 Whatman® FTA® punches in a 96-well PCR plate. The samples were incubated
321 at 50°C for an hour followed by protease inactivation by heating the samples
322 to 75°C for 30 min.
323 4. *ForensicGem* (FGM): Extraction using the ForensicGEM Universal DNA
324 extraction kit (cat: FUN0100; ZyGem) was performed according to the
325 manufacturer's guidelines.
326 5. *PicoPure* (PIP): Extraction using the ARCTURUS® PicoPure® DNA Extraction Kit
327 (cat: KIT0103; ThermoFisher) was performed according to the manufacturer's
328 guidelines, using 75 µL of the extraction buffer.

329

330 When the extracted DNA from all the above methods was present in a volume
331 greater than 25 µL, samples were cleaned with Agencourt AMPure XP beads
332 (Beckman-Coulter) and eluted in 25 µL nuclease-free H_2O. The entire DNA samples
333 were used downstream to make sequencing libraries. A summary of the number of
334 species, life stages, and conditions tested, is presented in **Table 1**.

335

336

337

338 **Table 1.** Summary of the number of species, life stages, and conditions tested

Species	Life stage	Extraction kit	Sample number
Ancylostoma caninum	Eggs	CGP	8
Ancylostoma caninum	L1	CGP	8
Ancylostoma caninum	L3	CGP	8
Ascaridia dissimilis	Eggs	CGP	8
Dirofilaria immitis	microfilaria	CGP	8
Dracunculus medinensis	L1	Nexttec	129
Haemonchus contortus	Egg	Bloodspot	10
Haemonchus contortus	Egg	CGP	10
Haemonchus contortus	Egg	ForensicGem	4
Haemonchus contortus	Egg	Nexttec	10
Haemonchus contortus	Egg	Picopure	10
Haemonchus contortus	Frozen Egg	Bloodspot	6
Haemonchus contortus	Frozen Egg	CGP	6
Haemonchus contortus	Frozen Egg	Nexttec	6
Haemonchus contortus	Frozen Egg	Picopure	6
Haemonchus contortus	L1	Bloodspot	10
Haemonchus contortus	L1	CGP	10
Haemonchus contortus	L1	ForensicGem	4
Haemonchus contortus	L1	Nexttec	10
Haemonchus contortus	L1	Picopure	10
Schistosoma mansoni	miracidia	Bloodspot	59
Schistosoma mansoni	miracidia	CGP	36
Schistosoma mansoni	miracidia	ForensicGEM	10
Schistosoma mansoni	miracidia	Nexttec	51

Schistosoma mansoni	miracidia	PicoPure	12
Strongyloides stercoralis	FL	CGP	8
Strongyloides stercoralis	iL3	CGP	8
Strongyloides stercoralis	L1	CGP	8
Trichuris muris	bleached eggs	CGP	8
Trichuris muris	eggs	CGP	8
Trichuris muris	L1	CGP	8

Library preparation and Sequencing

DNA sequencing libraries for all samples were prepared using a protocol designed for library preparation of laser capture microdissected biopsy (LCMB) samples using the Ultra II FS enzyme (New England Biolabs) for DNA fragmentation as previously described (Lee-Six et al., 2019). A total of 12 cycles of PCR were used (unless otherwise stated in **Table S2**) to amplify libraries and to add a unique 8-base index sequence for sample multiplexing.

Multiplexed libraries were sequenced using the Illumina MiSeq system with V2 chemistry 150 bp paired end (PE) reads. The *D. medinensis* samples were sequenced as part of a different study using the HiSeq 2500 with V4 chemistry 125 bp PE reads. In general, we performed sufficient low coverage sequencing on each sample to enable us to identify: (i) the proportion of on-target mapped reads, (ii) the proportion of duplicate reads, i.e. library artefacts, and (iii) the proportion of off-target contaminant reads.

Metadata for each sample, including sample IDs, sequencing lane IDs, ENA sample accession numbers, and data generated are described in **Table S2**. Raw sequence data will be made available under ENA study ID ERP114942.

Analysis

363 Reference genomes from each of the test species were obtained from WormBase
364 Parasite (Howe et al., 2017) Release 12. Raw sequence data for each species were
365 mapped to their respective reference genome using BWA-MEM (Li, 2013), after
366 which duplicate reads were marked using Picard (v2.5.0;
367 https://github.com/broadinstitute/picard). Samtools flagstats and bamtools stats
368 were used to characterise the outcome of the mapping, the results of which were
369 collated using MultiQC (Ewels et al., 2016). Data were manipulated and visualised in
370 the R (v3.5.0) environment using the following packages: ggplot2
371 (https://ggplot2.tidyverse.org/), patchwork
372 (https://github.com/thomasp85/patchwork), and dplyr
373 (https://dplyr.tidyverse.org/).
374
375 The code to reproduce the analysis and figures for this manuscript is described in
376 https://github.com/stephenrdoyle/helminth_extraction_wgs_test.
377
378

379 **Results & Discussion**

380 The aim of this work was to determine the feasibility and efficiency of using a low
381 input DNA extraction and library preparation approach for whole-genome
382 sequencing of individual helminth parasites. We have targeted immature life stages
383 that are found in the environment for which it is possible to collect samples non-
384 invasively. We first tested our approaches on the nematode *H. contortus* and the
385 trematode *S. mansoni*. Five approaches were tested using single egg (fresh and
386 frozen) and L3 of *H. contortus*, and miracidia of *S. mansoni.* We determined the
387 success of the library preparation approach by comparing the proportion of reads
388 mapped (**Figure 1 & 2**; top plots), representing on-target mapping as a measure of
389 specificity, as well as the proportion of duplicate reads (**Figure 1 & 2**; bottom plots),
390 which typically represent library preparation artefacts due to over-amplification of
391 DNA during PCR. Finally, we determined the proportion of reads associated with off-
392 target contamination by comparison of raw reads to a kraken contamination
393 database (**Table S2**; kraken_unassigned).

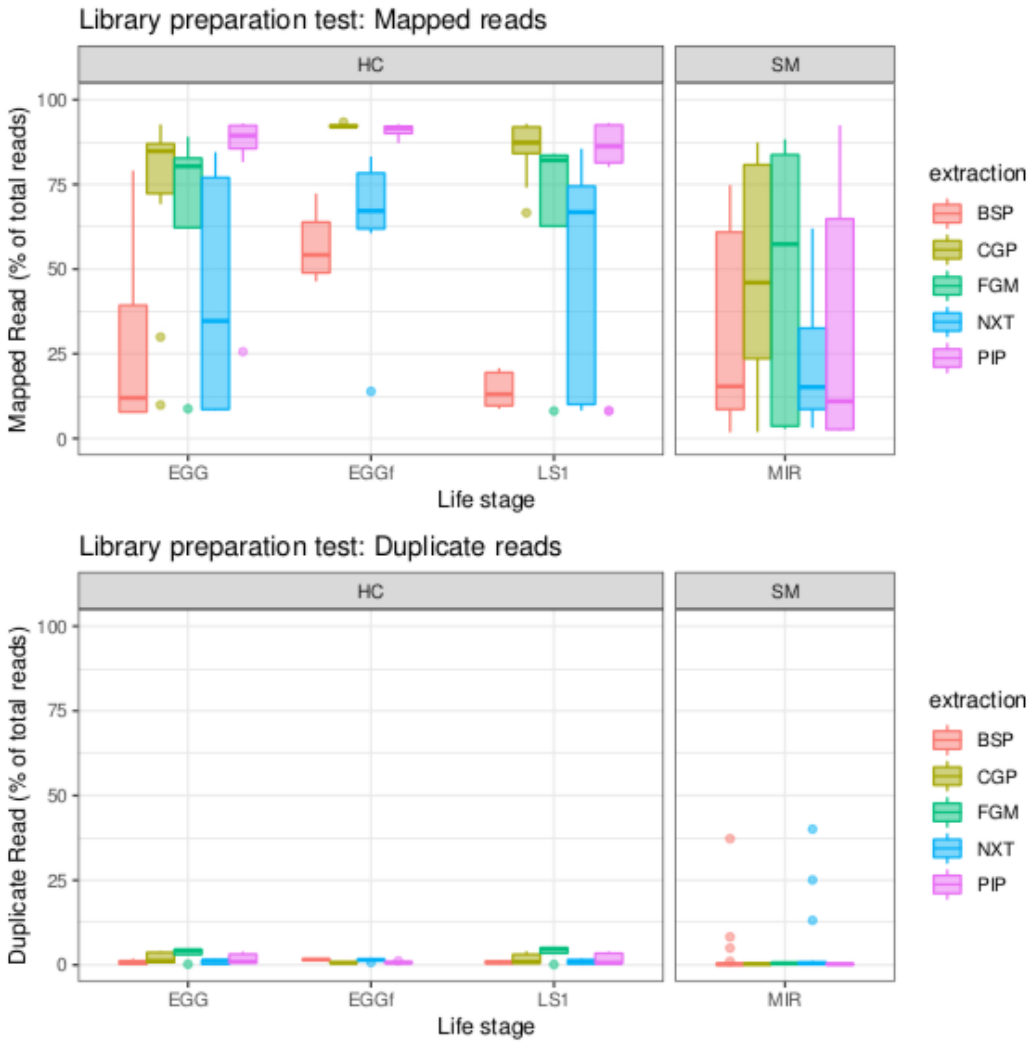
394

395 There were distinct differences in overall mapping between *H. contortus* and *S.*
396 *mansoni*; while some of this variation may reflect differences in extracting DNA from
397 different species, the *S. mansoni* samples generally had a greater proportion of
398 contaminating DNA present relative to the *H. contortus* samples (**Table S2**;
399 kraken_unassigned). We achieved some on-target mapping under all optimisation
400 conditions tested, however, significant variation was observed between approaches
401 (**Figure 1**). BSP and NXT generally performed poorly, with either low (median =
402 19.42; median absolute deviation (MAD) = 16.50) or significant variance (median =
403 66.20; MAD = 26.18) in mapping frequency between samples observed, respectively.
404 PIP performed consistently well with high mapping rates across all stages in *H.*
405 *contortus* (median = 90.30; MAD = 3.96), however, it was poor in *S. mansoni* (median
406 = 11.00; MAD = 12.73). The duplication rate of all conditions were within an
407 acceptable low range (median = 0.46, MAD = 0.4), with only 2% of samples having
408 greater than 5% duplicate reads. CGP and FGM performed most consistently
409 between stages and species; FGM had higher variance and duplication rates relative
410 to CGP across all samples tested, and therefore, CGP was chosen to explore further.

411

412

413



414

415 **Figure 1. Comparison of library preparation approaches on *Haemonchus contortus***
416 **eggs (EGG), frozen eggs (EGGf), and L1 (LS1) stages, and *Schistosoma mansoni***
417 **miracidia (MIR).** The top plot presents the percentage of reads that mapped to each
418 reference genome, and the bottom plot presents the proportion of duplicate reads
419 identified. Boxplots are coloured by the library preparation approach used. The
420 number of samples in each comparison is presented in Table 1.

421

422

423 We expanded our analysis to a total of eight helminth species for which samples
424 were available, including a total of six distinct life stages (**Figure 2**). All tested
425 samples were extracted with the CGP protocol, except for *D. medinensis*, which were

426 extracted using NXT and included for comparison. High variability in mapping was
427 observed between species, with 50% or greater mapping frequency achieved in at
428 least one life stage of 6 of the 8 species tested. Clear differences were observed
429 between multiple life stages tested within a species, likely reflecting differences in
430 extraction efficiency, for example: for *A. canium,* reads from eggs (median = 54.98)
431 mapped much more effectively than L1 (median = 3.61) or L3 stages (median = 7.57),
432 and in *S. stercoralis,* full-length females (median = 67.84) and L1 (median = 47.19)
433 performed better than L3 (median = 9.01). Interestingly, L1 larvae (median = 41.85)
434 and bleached eggs (median = 51.35) performed much better than untreated eggs
435 (median = 2.72) of *T. muris*; bleaching is experimentally used for promoting hatching
436 of *T. muris*, by dissolving the egg shell layers and in turn improving access to DNA
437 within. However, bleached eggs were embryonated and developmentally more
438 advanced than untreated unembryonated eggs, and therefore would have more
439 DNA available for library preparation. Similar to untreated *T. muris* eggs, *A. dissimilis*
440 eggs performed poorly (median = 0.16); for both species, few if any nematode
441 sequences were recovered and the majority of sequencing reads were
442 contaminating bacterial-derived contaminants, perhaps indicative of the challenge of
443 accessing material with the environmentally-resistant egg. Analysis of *A. dissimilis*
444 was further limited by the lack of a reference genome for this species; we mapped
445 against the available genome of *Ascaris lumbricoides*, the nearest species for which a
446 reference was available, and therefore at best, would have expected suboptimal
447 mapping due to sequence divergence. The duplication rates remain low for all
448 species and conditions tested, apart from *D. medinensis*, which was much higher
449 (median = 36.90). This was unexpected, given NXT did not produce high duplication
450 rates under the initial conditions tested with *H. contortus* or *S. mansoni*, nor were
451 there excessive PCR cycles used in this instance.
452

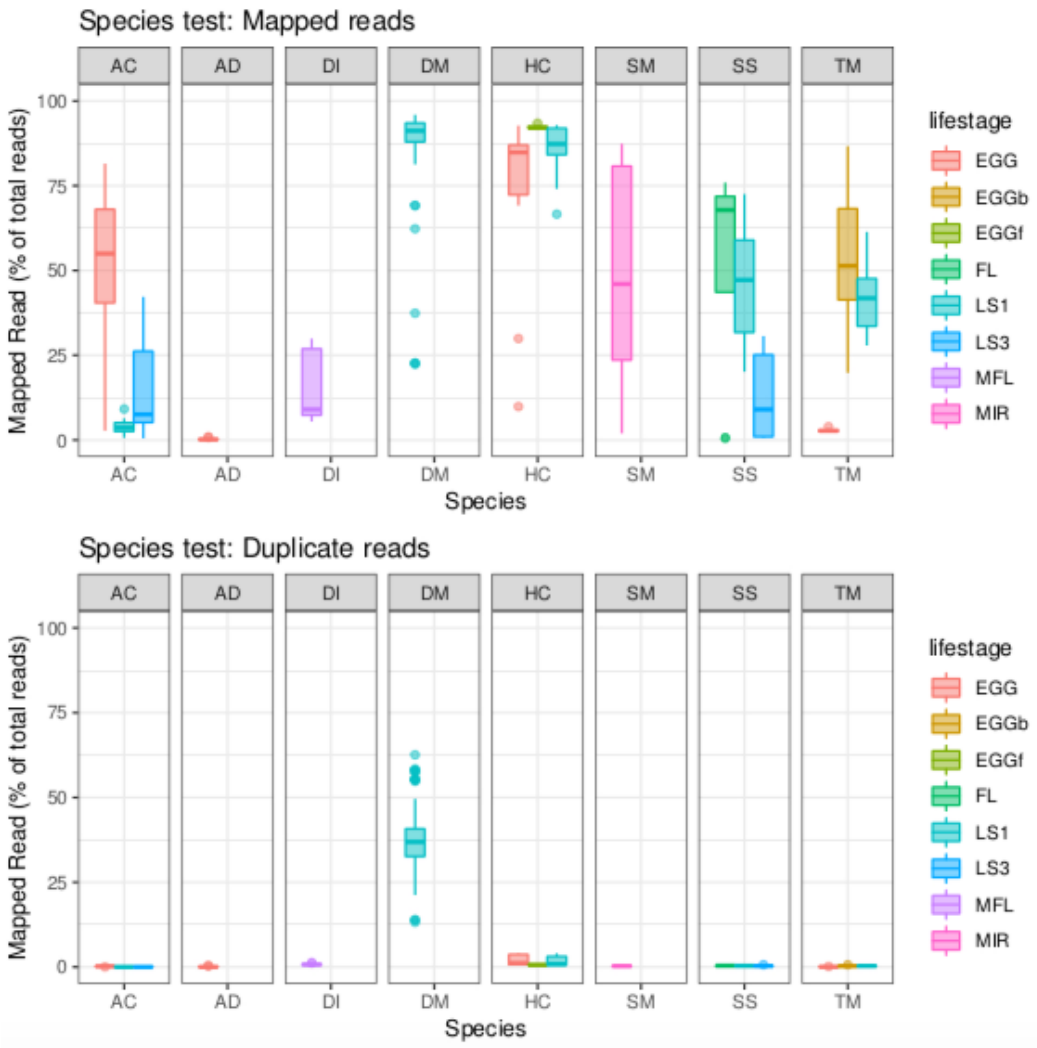


Figure 2. Comparison of library preparation from multiple life stages of 8 helminth species. All libraries were prepared with the CGP approach, except for *D. medinensis*, which was prepared with NXT and presented alongside for comparison. As in Figure 1, the top plot presents the percentage of reads that mapped to each reference genome, and the bottom plot presents the proportion of duplicate reads identified. Boxplots are coloured by the life stage assayed. Egg = untreated egg; EGGb = bleached egg; EGGf = frozen egg; FL = full-length female; LS1 = larval stage 1; LS3 = larval stage 3; MFL = microfilaria; MIR = miracidia. The number of samples in each comparison is presented in Table 1.

In summary, we present successful DNA extraction followed by whole-genome sequencing of individual parasites from 6 of 8 species examined. These results significantly extend the possibility of genomic analyses for life stages for which, at

468 best, were limited to low-resolution, low-throughput PCR based assays without the
469 addition of whole genome amplification. Whatman® FTA® cards provide a
470 convenient substrate for sample collection and storage, and do not limit the
471 application of direct DNA extraction and whole genome sequencing of parasite
472 samples, even for field samples as demonstrated for *S. mansoni* miracidia that were
473 collected and processed in Uganda before they were transported to the UK. Further
474 optimisation is required to improve the DNA recovery from eggs, for example, from
475 *A. dissimilis* and *T. muris*, to provide greater applicability of our approaches to
476 species that generate particularly environmentally-resistant stages, such as the soil
477 transmitted helminths. The application of whole genome sequencing to diagnose
478 and monitor helminth infections at scale is largely limited by the costs of library
479 preparation and sequencing, and therefore, will be restricted to niche applications of
480 the technology. However, targeting the mitochondrial genome, for example, by
481 whole genome sequencing may be a viable and cost-effective alternative, potentially
482 providing greater diagnostic information than low throughput PCR-based diagnostics
483 (**Table 2**). Continued development of genomic technologies and the associated
484 reduction in sequencing and library preparation costs will make screening large
485 samples by genome sequencing more routine as in viral (Dudas et al., 2017) and
486 bacterial (Domman et al., 2018) population studies. In doing so, the ability to derive
487 information-rich data for diagnostic and surveillance purposes using genomics
488 (Cotton et al., 2018) will be particularly informative as efforts to control human
489 infective helminths using MDA move from control to elimination.
490
491
492
493
494
495
496
497
498
499
500
501

502

503 **Table 2. Breakdown of sequencing strategies per species based on whole genome**
504 **sequencing at 30X coverage and whole genome sequencing to achieve 100X whole**
505 **mitochondrial genome coverage**

Species / stage	Genome assembly size [1]	Number of samples that can be multiplexed [2]	mtDNA/ nuclear ratio [3]	Sampled multiplexed targeting mtDNA at 100X coverage [2]	Nuclear genome coverage per 100X mtDNA genome
S. mansoni miracidia	409	173	142.77	7418	0.70
H. contortus egg	283	250	70.76	5313	1.41
H. contortus L1	283	250	44.55	3345	2.24
D. medinensis L1	103	688	70.81	14608	1.41
D. immitis microfilaria	88	805	14.27	3446	7.01
S. stercoralis FL	42	1687	11.06	5597	9.04

506

507 1. Genome size: Obtained from Wormbase Parasite v12
508 2. Multiplex: total number of samples per NovaSeq 6000 sequencing run with S4 2x150 bp PE chemistry
509 generating 2.5 Tb of data and accounting for 85% mapping rate. Theoretically possible, but may be
510 limited by barcoding available.
511 3. mtDNA/nuclear ratio: based on coverage of mitochondrial and nuclear derived sequencing reads,
512 normalised to the nuclear genome coverage

513

514

Acknowledgments

Work performed at the Wellcome Sanger Institute is supported by Wellcome Trust (grant 206194) and by the Biotechnology and Biological Sciences Research Council (BB/M003949/1). We thank Alison Elliot for access to *S. mansoni* samples collected from Uganda, the collection of which was supported by Wellcome Trust (grant 095778/Z/11/Z). *S. mansoni* samples were also obtained from the Schistosomiasis Collection at the Natural History Muesem (NHM) [SCAN], which is funded with support from the Wellcome Trust (grant no. 104958/Z/14/Z). We thank the Carter Center for supporting molecular work on *D. medinensis,* and the Guinea worm eradication program for making samples available.

Author contributions

Conceptualization: SRD, JAC, NH

Methodology: SRD, GS, NH

Software: SRD

Formal analysis: SRD

Investigation: SD, GS

Resources: FA, TJ, JL, JBC, PJC, JPW, TC, ET, TC, MAD-C, PE, RL, KM, CM, TM, BS, PTO, JVT, AW, ED, RK, DB, MB

Data Curation: SRD, NH

Writing – original draft preparation: SRD

Writing – review and editing: All authors

Visualisation: SRD

Supervision: SRD, JAC, NH

Project administration: NH

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Data

Table S1 - Summary of total samples analysed by species, life stage, and DNA extraction protocol (excel workbook)

Table S2 - Complete metadata per sample (excel workbook)

548 **References**

549 Avramenko, R. W., Redman, E. M., Lewis, R., Yazwinski, T. A., Wasmuth, J. D., and Gilleard, J.
550 S. (2015). Exploring the Gastrointestinal "Nemabiome": Deep Amplicon Sequencing to
551 Quantify the Species Composition of Parasitic Nematode Communities. *PLoS One* 10,
552 e0143559. doi:10.1371/journal.pone.0143559.

553 Avramenko, R. W., Redman, E. M., Melville, L., Bartley, Y., Wit, J., Queiroz, C., et al. (2019).
554 Deep amplicon sequencing as a powerful new tool to screen for sequence
555 polymorphisms associated with anthelmintic resistance in parasitic nematode
556 populations. *International Journal for Parasitology* 49, 13–26.
557 doi:10.1016/j.ijpara.2018.10.005.

558 Boué, F., El Berbri, I., Hormaz, V., Boucher, J.-M., El Mamy, A. B., Traore, A., et al. (2017). Use
559 of FTA® card methodology for sampling and molecular characterization of Echinococcus
560 granulosus sensu lato in Africa. *Exp. Parasitol.* 173, 29–33.

561 Campbell, S. J., Stothard, J. R., O'Halloran, F., Sankey, D., Durant, T., Ombede, D. E., et al.
562 (2017). Urogenital schistosomiasis and soil-transmitted helminthiasis (STH) in
563 Cameroon: An epidemiological update at Barombi Mbo and Barombi Kotto crater lakes
564 assessing prospects for intensified control interventions. *Infect Dis Poverty* 6, 49.

565 Cotton, J. A., Berriman, M., Dalén, L., and Barnes, I. (2018). Eradication genomics-lessons for
566 parasite control. *Science* 361, 130–131.

567 Crellen, T., Walker, M., Lamberton, P. H. L., Kabatereine, N. B., Tukahebwa, E. M., Cotton, J.
568 A., et al. (2016). Reduced Efficacy of Praziquantel Against Schistosoma mansoni Is
569 Associated With Multiple Rounds of Mass Drug Administration. *Clin. Infect. Dis.* 63,
570 1151–1159.

571 Domman, D., Chowdhury, F., Khan, A. I., Dorman, M. J., Mutreja, A., Uddin, M. I., et al.
572 (2018). Defining endemic cholera at three levels of spatiotemporal resolution within
573 Bangladesh. *Nat. Genet.* 50, 951–955.

574 Doyle, S. R., Laing, R., Bartley, D. J., Britton, C., Chaudhry, U., Gilleard, J. S., et al. (2018). A
575 Genome Resequencing-Based Genetic Map Reveals the Recombination Landscape of an
576 Outbred Parasitic Nematode in the Presence of Polyploidy and Polyandry. *Genome Biol.*
577 *Evol.* 10, 396–409.

578 Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., Faria, N. R., et al. (2017). Virus
579 genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* 544,
580 309–315.

581 Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis
582 results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048.

583 Gower, C. M., Shrivastava, J., Lamberton, P. H. L., Rollinson, D., Webster, B. L., Emery, A., et
584 al. (2007). Development and application of an ethically and epidemiologically
585 advantageous assay for the multi-locus microsatellite analysis of Schistosoma mansoni.
586 *Parasitology* 134, 523–536.

587 Howe, K. L., Bolt, B. J., Shafie, M., Kersey, P., and Berriman, M. (2017). WormBase ParaSite -
588 a comprehensive resource for helminth genomics. *Mol. Biochem. Parasitol.* 215, 2–10.

589 Howick, V. M., Russell, A., Andrews, T., Heaton, H., Reid, A. J., Natarajan, K. N., et al. (2019).
590 The Malaria Cell Atlas: a comprehensive reference of single parasite transcriptomes
591 across the complete Plasmodium life cycle. *bioRxiv*, 527556. doi:10.1101/527556.

592 Jaleta, T. G., Zhou, S., Bemm, F. M., Schär, F., Khieu, V., Muth, S., et al. (2017). Different but
593 overlapping populations of Strongyloides stercoralis in dogs and humans-Dogs as a
594 possible source for zoonotic strongyloidiasis. *PLoS Negl. Trop. Dis.* 11, e0005752.

595 Le Clec'h, W., Chevalier, F. D., McDew-White, M., Allan, F., Webster, B. L., Gouvras, A. N., et
596 al. (2018). Whole genome amplification and exome sequencing of archived schistosome
597 miracidia. *Parasitology* 145, 1739–1747.

598 Lee-Six, H., Ellis, P., Osborne, R. J., Sanders, M. A., Moore, L., Georgakopoulos, N., et al.
599 (2019). The landscape of somatic mutation in normal colorectal epithelial cells. *bioRxiv*.
600 doi:10.1101/416800.

601 Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-
602 MEM. *arXiv [q-bio.GN]*. Available at: http://arxiv.org/abs/1303.3997.

603 Lok, J. B. (2007). Strongyloides stercoralis: a model for translational research on parasitic
604 nematode biology. *WormBook*, 1–18.

605 Maclean, M. J., Savadelis, M. D., Coates, R., Dzimianski, M. T., Jones, C., Benbow, C., et al.
606 (2017). Does evaluation of in vitro microfilarial motility reflect the resistance status of
607 Dirofilaria immitis isolates to macrocyclic lactones? *Parasit. Vectors* 10, 480.

608 Marek, M., Zouhar, M., Douda, O., Maňasová, M., and Ryšánek, P. (2014). Exploitation of
609 FTA cartridges for the sampling, long-term storage, and DNA-based analyses of plant-
610 parasitic nematodes. *Phytopathology* 104, 306–312.

611 Moore, L., Leongamornlert, D., Coorens, T. H. H., Sanders, M. A., Ellis, P., Dawson, K., et al.
612 (2018). The mutational landscape of normal human endometrial epithelium. *bioRxiv*.
613 doi:10.1101/505685.

614 Ngara, M., Palmkvist, M., Sagasser, S., Hjelmqvist, D., Björklund, Å. K., Wahlgren, M., et al.
615 (2018). Exploring parasite heterogeneity using single-cell RNA-seq reveals a gene
616 signature among sexual stage Plasmodium falciparum parasites. *Exp. Cell Res.* 371, 130–
617 138.

618 Platt, R. N., McDew-White, M., Le Clec'h, W., Chevalier, F. D., Allan, F., Emery, A. M., et al.
619 (2019). Ancient hybridization and introgression of an invadolysin gene in schistosome
620 parasites. *bioRxiv*. doi:10.1101/539353.

621 Reid, A. J., Talman, A. M., Bennett, H. M., Gomes, A. R., Sanders, M. J., Illingworth, C. J. R., et
622 al. (2018). Single-cell RNA-seq reveals hidden transcriptional variation in malaria
623 parasites. *Elife* 7. doi:10.7554/eLife.33105.

624 Sabina, J., and Leamon, J. H. (2015). Bias in Whole Genome Amplification: Causes and
625 Considerations. *Methods Mol. Biol.* 1347, 15–41.

626 Sanya, R. E., Nkurunungi, G., Hoek Spaans, R., Nampijja, M., O'Hara, G., Kizindo, R., et al.
627 (2018). The impact of intensive versus standard anthelminthic treatment on allergy-
628 related outcomes, helminth infection intensity and helminth-related morbidity in Lake
629 Victoria fishing communities, Uganda: results from the LaVIISWA cluster randomised

630 trial. *Clin. Infect. Dis.* doi:10.1093/cid/ciy761.

631 Shortt, J. A., Card, D. C., Schield, D. R., Liu, Y., Zhong, B., Castoe, T. A., et al. (2017). Whole
632 Genome Amplification and Reduced-Representation Genome Sequencing of
633 Schistosoma japonicum Miracidia. *PLoS Negl. Trop. Dis.* 11, e0005292.

634 Small, S. T., Labbé, F., Coulibaly, Y. I., Nutman, T. B., King, C. L., Serre, D., et al. (2018).
635 Human Migration and the Spread of the Nematode Parasite Wuchereria bancrofti.
636 *bioRxiv*. doi:10.1101/421248.

637 Sulston, J., and Hodgkin, J. (1988). "Methods," in *The Nematode Caenorhabditis elegans*, ed.
638 W. B. Wood (Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press).

639 Trevino, S. G., Nkhoma, S. C., Nair, S., Daniel, B. J., Moncada, K., Khoswe, S., et al. (2017).
640 High-Resolution Single-Cell Sequencing of Malaria Parasites. *Genome Biol. Evol.* 9,
641 3373–3383.

642 Tsai, I. J., Hunt, M., Holroyd, N., Huckvale, T., Berriman, M., and Kikuchi, T. (2014).
643 Summarizing specific profiles in Illumina sequencing from whole-genome amplified
644 DNA. *DNA Res.* 21, 243–254.

645 Wakelin, D. (1967). Acquired immunity to Trichuris muris in the albino laboratory mouse.
646 *Parasitology* 57, 515–524.

647 Webster, B. L. (2009). Isolation and preservation of schistosome eggs and larvae in
648 RNAlater® facilitates genetic profiling of individuals. *Parasit. Vectors* 2, 50.

649 Webster, B. L., Emery, A. M., Webster, J. P., Gouvras, A., Garba, A., Diaw, O., et al. (2012).
650 Genetic diversity within Schistosoma haematobium: DNA barcoding reveals two distinct
651 groups. *PLoS Negl. Trop. Dis.* 6, e1882.

652 Xiao, N., Remais, J. V., Brindley, P. J., Qiu, D.-C., Carlton, E. J., Li, R.-Z., et al. (2013).
653 Approaches to genotyping individual miracidia of Schistosoma japonicum. *Parasitol.*
654 *Res.* 112, 3991–3999.

655