

Population genomics and epidemiology of *Schistosoma mansoni*

Thomas Joseph Crellen

Department of Infectious Disease Epidemiology

School of Public Health

Imperial College London

PhD Thesis

Declaration of Originality

All work presented in this thesis is the candidate's own, under the supervision of Prof. Joanne Webster, Dr. James Cotton and Dr. David Aanensen. Any work by others has been appropriately referenced.

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Abstract

Schistosoma mansoni is a metazoan parasite that infects millions of people across the developing world. I applied coalescence models to genome-wide single nucleotide polymorphisms (SNPs) from ten global isolates of *S. mansoni* and an isolate of the closely related *Schistosoma rodhaini* to infer the dates of: i) the speciation between *S. mansoni* and *S. rodhaini*; ii) the spread of *S. mansoni* across sub-Saharan Africa; and iii) the migration to the New World. Further, I used this sequence data to document signatures of directional selection between *S. mansoni* and *S. rodhaini* to uncover genes involved in host adaptation and under balancing selection in *S. mansoni*.

I conducted a field study in Uganda to collect epidemiological data from primary schools that had been subject to contrasting intensities and durations of Mass Drug Administration (MDA). Children in endemic areas were examined for *S. mansoni* up to three days before and after treatment with praziquantel. I fitted a mixed-effects model to the egg count data and found that individuals from schools that were treated most intensively had a lower estimated mean egg reduction rate compared with individuals from less intensively treated schools, suggesting that more rounds of MDA may be resulting in *S. mansoni* parasites that are better tolerating anthelmintic treatment.

Finally, I collected *S. mansoni* miracidia from patients in Uganda and developed a novel method for generating whole-genome sequences. In total, 213 miracidia from 35 patients were sequenced. Using a dataset of 5.4 million SNPs, I documented the population structure, characterised regions of the genome under selection and associated SNPs against the clearance phenotype. I uncovered the strongest signal of selection and the highest association with host clearance phenotype in a region at the beginning of chromosome 2 that contains two calcium channels; variants in this region may therefore be underlying the observed reduction in praziquantel efficacy.

Table of Contents

List of Tables	5
List of Figures	6
Acknowledgements	8
Chapter 1	9
Historical Background.....	9
Taxonomic background.....	10
Schistosome lifecycle.....	11
Pathology	13
Global Distribution and Epidemiology.....	14
Diagnostics.....	16
Treatment and Control	17
Praziquantel	19
Drug resistance in praziquantel?.....	19
Past population genetic studies of <i>S. mansoni</i>	21
Genetic population structure of <i>S. mansoni</i> populations	23
Effect of praziquantel on <i>S. mansoni</i> populations.....	24
Next Generation Sequencing and Parasite Genomics	26
Exploring drug resistance with genomics.....	28
Thesis outline.....	31
Chapter 2	33
Abstract.....	33
Introduction	34
Methods & Materials.....	36
Results	40
Discussion	62
Chapter 3	69
Abstract.....	69
Introduction	70
Methods & Materials.....	71
Results	79
Discussion	90
Ethical Statement	92
Chapter 4	93
Abstract.....	93
Introduction	94
Methods and Materials	97
Results	116
Discussion	138
Chapter 5	143
References.....	149
Appendix 1	169
Appendix 2	170
Appendix 3	171
Appendix 4	172

List of Tables

Table 1. Global isolates of <i>S. mansoni</i> sequenced and number of variants.....	42
Table 2. Effective population size estimates obtained from G-PhoCS.....	58
Table 3. Sample estimates of the prevalence and intensity of <i>Schistosoma mansoni</i> infection in children from six primary schools in eastern Uganda before treatment with praziquantel. ...	75
Table 4. Estimated coefficient posterior distributions of the generalized linear mixed model fitted to <i>Schistosoma mansoni</i> egg counts collected from 414 children from six primary schools in eastern Uganda before and after treatment with praziquantel.....	81
Table 5. Estimated mean egg reduction rates (ERRs) of <i>Schistosoma mansoni</i> in 414 children from six primary schools in eastern Uganda following treatment with praziquantel.	86
Table 6. Estimated mean egg reduction rates (ERRs) of <i>Schistosoma mansoni</i> after resampling by Kato-Katz readings.....	87
Table 7. Sampled hosts ($n=35$) and miracidia ($n=190$) mean, range and total by school and time point.....	110
Table 8. Single nucleotide polymorphisms and samples present in the dataset after applying filters from the Genome Analysis Toolkit (GATK) and PLINK.	110
Table 9. Mean F_{ST} values at different spatial scales. Values are calculated using 2.11 millions SNPs from 190 <i>S. mansoni</i> miracidia collected in Uganda at the level of individual individuals, schools and districts.	124
Table 10. Nucleotide diversity values by chromosome and time point from miracidia collected in Uganda before ($n=54$) and after ($n=49$) treatment with praziquantel.....	125
Table 11. Genes from a 256 kb region of selection on chromosome 2 of the V6 <i>S. mansoni</i> reference genome. Coordinates indicate the location of the gene in kilobases.....	134
Table 12. Non-synonymous variants and indels within the exons of calcium channel genes inside the 256 kb window of selection on chromosome 2.	136

List of Figures

Figure 1. Number of SNPs when samples called against <i>S. mansoni</i> reference V5.2 by calling algorithm.....	43
Figure 2. Maximum likelihood phylogeny of <i>Schistosoma</i> isolates based on whole-genome SNP data.....	45
Figure 3. Principal component analyses of global isolate SNP data.	46
Figure 4. Histogram showing d_N/d_S values for 9514 coding regions in <i>S. mansoni</i> when compared against <i>S. rodhaini</i> using a branch model in codeml.	50
Figure 5. Population genetic evidence for balancing and directional selection on <i>S. mansoni</i> genes.	51
Figure 6. Estimate of the mutation rate (per basepair per generation) of <i>S. mansoni</i>	55
Figure 7. A measure of linkage disequilibrium (r^2) across the <i>S. mansoni</i> genome, calculated from phased SNP calls.....	56
Figure 8. Models used for G-PhoCS coalescence simulations.	57
Figure 9. Date of <i>S. mansoni</i> – <i>S. rodhaini</i> divergence and archaeological evidence for fishing in Africa.	59
Figure 10. PSMC estimates of effective population size variation through time.....	60
Figure 11. Bootstrap confidence intervals (100 replicates) around PSMC estimates.....	61
Figure 12. Summary of <i>Schistosoma mansoni</i> population history based on the evidence from this chapter.....	66
Figure 13. Maps of study sites in Uganda.	72
Figure 14. Photos taken during recruitment in Mayuge District, Uganda.	73
Figure 15. Flow diagram of patient recruitment and exclusion criteria for egg reduction rate analysis after treatment with praziquantel (PZQ) and albendazole (ALB).	77
Figure 16. Estimated egg reduction rates (ERRs) of <i>Schistosoma mansoni</i> from 414 children in eastern Uganda treated with praziquantel.....	82
Figure 17. Estimated egg counts of <i>Schistosoma mansoni</i> infections in 414 children in eastern Uganda before (BT) and after (AT) treatment with praziquantel.....	84
Figure 18. Posterior distribution of average egg reduction rates (ERRs) of <i>Schistosoma mansoni</i> in 414 children from 6 primary schools in eastern Uganda treated with praziquantel.	89
Figure 19. Technician Andrina Nankasi collecting miracidia during field studies in eastern Uganda.	100

Figure 20. Boxplots comparing percentage of reads mapping to <i>S. mansoni</i> v5.2 reference genome by amplification method.....	102
Figure 21. Plot showing sequencing depth (median 11x) for sample Bw9_D along 7 autosomal chromosomes and ZW sex chromosome.....	106
Figure 22. Comparison of sample and genotype missingness, and minor allele frequency using 10.61 million variants and 213 miracidia.....	109
Figure 23. Number of miracidia ($n=190$) per host ($n=35$) by school and time point.....	111
Figure 24. Principal components analysis of 5.31 million filtered SNPs from 190 <i>S. mansoni</i> miracidia collected in Uganda.....	119
Figure 25. Principal components analysis using 15.5 million SNPs from 190 Ugandan miracidia along with isolates Lake-Albert, Lake-Victoria and Cameroon sequenced from adult worms in Chapter 2.....	120
Figure 26. Linkage disequilibrium as measured by r^2 between pairs of genotypes within chromosomes.....	123
Figure 27. Values of nucleotide diversity of <i>S. mansoni</i> miracidia (π ; π) for each chromosome...	126
Figure 28. Values of F_{ST} comparing miracidia from before ($n=54$) and after ($n=49$) treatment with praziquantel from hosts in three intensively treated schools in Uganda.....	127
Figure 29. Values of F_{ST} plotted along the <i>S. mansoni</i> genome from 190 miracidia. The subpopulations are within hosts ($n=35$).....	131
Figure 30. The integrated haplotype score (iHS) plotted along the 7 autosomal chromosomes of the <i>S. mansoni</i> genome from 190 miracidia.	132
Figure 31. Case control association study on the top 1% of SNPs under selection in 190 <i>S. mansoni</i> miracidia after removing markers in close linkage (6698 variants).....	133
Figure 32. Expression data from RNAseq of juvenile (schistosomulae) and adult <i>S. mansoni</i> worms from an <i>in vivo</i> model from different time points (in days) post-infection.	137

Acknowledgements

I would like to thank my supervisors Prof. Joanne Webster, Dr. James Cotton and Dr. David Aanensen for their tremendous support throughout my PhD and their advice and contributions that have shaped the thesis.

In addition I would like to thank Dr. Poppy Lamberton, formerly at Imperial College London, for her assistance in setting up fieldwork in Uganda along with the technicians, nurses and drivers from the Vector Control Division, Ugandan Ministry of Health, Mr. Moses Arinaitwe, Mr. Moses Adriko, Ms. Aida Wamboko, Ms. Annet Enzaru, Ms. Andrina Nankasi, Mr. Aaron Atuhairu and Mr. Fiddi Rugigana, whose hard work in data collection was instrumental to the study. I would also like to thank the teachers from the schools I visited for their assistance and cooperation along with officials in the Health and Education Departments in districts of Mayuge and Tororo, Uganda.

Much of the work in this thesis took place at the Wellcome Trust Sanger Institute, therefore I must thank the Parasite Genomics team for allowing me to work alongside them; in particular the group leader Dr. Matthew Berriman for his support and dedication to funding the sequencing aspects of this project. Ms. Nancy Holroyd was instrumental in coordinating the library preparation and sequencing of my samples. Ms Arporn “Koi” Wangwiwatsin of WTSI also kindly provided RNA-seq data used in Chapter 4.

The samples in Chapter 2 were selected by Dr. Bonnie Webster at the Natural History Museum and prepared by Dr. Fiona Allen at the NHM and Mr. Thomas Huckvale at the WTSI prior to sequencing. I was assisted in developing the protocol for DNA extraction/ amplification of samples for sequencing in Chapter 4 by Dr. Fiona Allen at the NHM and Dr. Hayley Bennett at the WTSI.

The statistical analysis in Chapter 3 was aided by Dr. Martin Walker of Imperial College London who also proof-read this Chapter. Dr Stephen Doyle of the WTSI also provided valuable comments on Chapter 4.

I am grateful to the MRC for funding my PhD Studentship along with the Schistosomiasis Control Initiative for supporting the costs of fieldwork and the WTSI’s core funding for reagents and sequencing.

Finally, I would like to thank family and friends for their support during my studies, in particular my housemates John Lees and Sumana Sharma for their encouragement during the tough times and celebrations in the better ones.

Chapter 1

Introduction

Historical Background

When schistosomes were first described by Theodor Bilharz in 1851, the field of biology was still a largely descriptive science. The century and a half that followed is littered with milestones in evolutionary theory [1], molecular biology [2], population genetics [3] and statistical modelling [4] that have made the present thesis possible. While this thesis can be considered in the field of ‘biological’ or ‘biomedical’ science, the majority of the analysis has been performed computationally. The fields of biology and computation would have seemed strange bedfellows in 1851, though during the second half of the Twentieth Century radical improvements to biological knowledge and the emergence of computer science have led to their present intertwining. In 1953, the same year that Crick and Watson elucidated the structure of DNA, scientists under John von Neumann in the Los Alamos Laboratory built the first random access memory computer to calculate the power of hydrogen bombs. The parallels between biochemical structures that encoded life as discrete bases and circuitry that encoded computation as binary bits were not lost on the researchers of the Manhattan Project [5]. Decoding the discrete information of life was made possible by the technological developments of Fred Sanger in the 1970s [6], though given the cost and low-throughput of early sequencing it took further technological advances before these methods could be applied to sequencing populations of organisms. Over the last ten years, sequencing technology has fallen in cost and increased in throughput at a rate greater than Moore’s Law [7]; the abundance of sequencing that has followed is transforming the landscape of biology and epidemiology into sciences driven by ‘big data’.

As the size of datasets has increased, improvements to computation have been necessary: both in hardware to process and store information and in methods, to improve the algorithms and statistical methods that make sense of it. Work early in the twentieth century by Ronald Fisher established much of the battery of statistical tests that are routinely applied to biological data (ANOVA, maximum likelihood, p -values) and crucially merged his statistical ideas with those on evolutionary theory in 1930 [8]. Subsequent developments of Fisher’s methods, such as the

generalised linear model, have been crucial tools in quantitative sciences such as epidemiology, in which studies often aim to infer associations in the context of multiple covariates [4]. Bayesian statistics also became more widely used in biology in order to incorporate prior information on the likelihood of outcomes [9]. In the late 1940s, physicists working with John von Neumann developed an algorithm that employed stochastic perturbations to explore the fates of a branching process [5]. Although the algorithm was first used to understand the collisions of neutrons in an atomic bomb, it was subsequently applied to Markov models and the algorithm (Markov Chain Monte Carlo; MCMC) underlies parameter estimation in many present-day Bayesian statistical methods, including those employed in this thesis. While these statistical methods were developed over fifty years ago, it is only recently that they have become computationally tractable for large datasets [10]. As researchers in the life sciences, we are currently situated at a point where statistical methods, computational power and massive data have converged to give us unprecedented resources to understand the natural world.

This thesis is focussed on the application of whole-genome sequencing to the human parasite *Schistosoma mansoni*, a causative agent of schistosomiasis, combined with statistical modelling of epidemiological data of the disease gathered from endemic settings; therefore a general discussion of the parasite's biology will be the starting point for the introduction. This will lead onto a discussion on epidemiology and disease control in endemic countries, past work on the population genetics of *S. mansoni*, the sequencing of the *S. mansoni* genome and prospective applications of parasite genomics in general.

Taxonomic background

The genus *Schistosoma* falls within the class Trematoda consisting of flukes; parasitic flatworms that are dorsoventrally flattened, bilaterally symmetrical and lack a body cavity. Flukes typically have a complex lifecycle that includes an invertebrate intermediate host and a vertebrate definitive host. Molluscs are the most common intermediate host for trematodes and the two phyla have shared a history of coevolution stretching back over millions of years [11]. Fish or birds are the most common definitive hosts of trematodes. As these vertebrates predate snails this suggests that trematodes may have adapted to survive in vertebrate hosts; predominantly in the intestines as this is where snails are ingested [12].

The family schistosomatidae are unique within Trematoda as all species are dioecious. While this development may have made it more challenging for schistosomes to complete their lifecycle, in that both male and female flukes are required to enter the definitive host and meet at the same site, it may have been driven by the need for physiological specialization as males are broader and more 'muscular' and thus are able to hold egg-laying females in place within blood vessels against the flow of circulation [13]. Of the twenty identified species within the genus *Schistosoma* [14], five are known to be human-infective (*S. mansoni*, *S. haematobium*, *S. intercalatum*, *S. japonicum*, *S. mekongi*) and their distributions span the tropical regions of Africa, Asia and South America [15]. Recently the geographic range of schistosomes has expanded to southern Europe as an outbreak led to *Schistosoma haematobium* and hybrid strains infecting at least 120 people in Corsica during the summer of 2013 [16]. Estimates of the number of people infected by the three main human-infective species (*S. mansoni*, *S. haematobium*, *S. japonicum*) carry large uncertainties, although conservative estimates suggest at least 230 million people have schistosomiasis worldwide, of which *S. mansoni* contributes an estimated 54 million infections in sub-Saharan Africa [15, 17]. With regards to *S. mansoni* specifically, the majority of cases are found in sub-Saharan Africa with a smaller number of infections present in Yemen, though this country is now close to elimination [18], and the New World; predominantly Brazil [19]. This global distribution that spans both the Old and New World is hypothesised to be due to the Atlantic Slave Trade, in which Africans transported to the New World carried *S. mansoni* parasites and established transmission of intestinal schistosomiasis in regions with compatible snails [20]. Genomic evidence to support this theory is presented in Chapter 2 of this thesis and the publication is presented in Appendix 3. Within endemic countries, areas of high prevalence and intensity are focussed around water bodies such as Lakes Victoria and Albert in East Africa [21] or water development projects such as irrigation schemes and dams [22].

Schistosome lifecycle

While the lifecycle of schistosomes varies by species, the developmental stages of adult worms, eggs, miracidia, sporocysts, cercariae, schistosomulae are common to all human-infective species. Pairs of adult male and female worms reside in the mesenteric veins draining the intestine in the case of intestinal schistosomiasis (*S. mansoni*, *S. intercalatum*, *S. japonicum*, *S. mekongi*), or

colonise the circulatory system surrounding the urogenital tract in the case of urogenital schistosomiasis (*S. haematobium*). The location of the adult worms indicates the mode of transmission by which eggs are released into the environment, which is through urine in *S. haematobium* and stool for other species. The diagnostic implications of this are that stool samples are required for the Kato-Katz thick smear that diagnoses intestinal schistosomiasis compared with urine filtration that diagnoses *S. haematobium*. The adult worms' location also impacts on pathology, whereby the majority of damage caused by *S. haematobium* eggs is to the urogenital tract whereas in other species chronic pathology principally affects the liver and spleen.

As this thesis focuses on *S. mansoni*, a more detailed discussion of the lifecycle and pathology will be restricted to this species. Laterally spined eggs are released in stool from infected individuals, and after only a few minutes of contact with fresh water hatch into free-swimming miracidia that live for 4-16 hours depending on conditions [23]. Upon reaching a compatible snail of the genus *Biomphalaria* spp., the miracidium attaches to the snail skin and, through the use of penetration glands and muscular squirming, penetrates the snail. Once inside, the ciliated surface of the miracidium used for motility is shed within two hours and a new tegument, a protective layer at the host-parasite interface, is formed [24]. The parasite thus develops into a primary sporocyst, which in turn develops germ-cell derived daughter sporocysts that produce thousands of single-sex cercariae around a month after the initial infection and are shed over a period of several weeks in field conditions [25]. Exposing the snail to light and heat triggers cercarial shedding. Once shed, cercariae swim towards light, follow a thermal gradient and respond to fatty acids on the surface of human skin to increase their likelihood of coming into contact with humans [26]. Once attached to their host, the cercariae lose their tails, thus transforming into schistosomulae. Proteases are required to penetrate the dermis, which is achieved by glands secreting an enzyme (cercarial elastase) that degrades insoluble elastin in skin [27]. The transformed schistosomula must then undergo migration to the final resting site of the portal veins. The details of this epic journey have largely been illuminated through the use of mouse models with the assumption that schistosomulae migration in humans follows a similar time course and trajectory. Briefly, schistosomulae take up to five days to migrate through the skin after penetration, whereby they enter into the capillaries. The parasites migrate into the lungs via the venous circulation. Schistosomes break out from the lungs through the pulmonary capillaries and are carried through the heart into systemic circulation and reach the portal system of the liver where they begin to

feed on blood cells and grow [28, 29]. By 25-30 days from first infection in the mammalian host the worms reach the mesenteric veins where they pair up, attain full maturity and the first fertilised eggs are laid. Indeed, pairing is required for the female schistosome to attain full maturity and if the male is decoupled she will regress in her development showing suboptimal growth and a reduced egg output [30].

Pathology

The pathology attributed to infection with *S. mansoni* can be considered as: a) acute pathology, known as Katayama syndrome, which is principally observed in travellers; and b) chronic pathology, also known as bilharzia or schistosomiasis, which is mainly experienced by individuals living in endemic areas as the pathology results from prolonged exposure to a high egg burden caused by multiple co-infecting worms.

Katayama syndrome is a systemic hypersensitivity reaction to migrating schistosomulae and egg deposition and occurs typically 14-84 days after the initial exposure to schistosomes in individuals with no acquired immunity [31]. An urticarial rash can appear several hours after cercariae have migrated through the skin and persist as lesions for several days, though dermatitis may also have a delayed onset of 1-12 weeks following heavy cercarial exposure. Other symptoms attributable to acute schistosomiasis are non-specific such as fever, fatigue, non-productive cough and eosinophilia [31]. The symptoms of Katayama fever are similar to 'Swimmer's Itch', which is a hypersensitivity reaction resulting in skin eruptions caused by the cercariae of animal-infective (often avian) schistosomes penetrating human skin. Cases of Swimmer's Itch have been reported in temperate latitudes as far north as the United Kingdom and Iceland. [32]

Pathology from chronic intestinal schistosomiasis occurs principally in the liver and spleen as a result of eggs swept into these organs by the portal system, rather than traversing the endothelium of the vein, and subsequently the epithelium of the intestine before being deposited in stool. The first description of this pathology came in 1904 from the English pathologist William St Clair Symmers who worked in the same Egyptian hospital as Theodore Bilharz. Commenting on a deceased patient, he wrote that

“To the naked eye the liver is somewhat enlarged, and the surface shows a peculiar shagreened appearance owing to a pronounced increase of the perivascular connective tissue. A small blood vessel may often be seen running up to such a mass of concentrically arranged young fibrous tissue, and lying in the centre is an ovum, as if it had become impacted in the vessel and had by its presence produced the proliferation of tissue which resulted in cirrhosis. This cirrhosis is, I believe, due to the presence of the Bilharzia ovum” [33, 34].

When schistosome eggs become lodged in tissue the granulomas that form around them develop through stages, characterised initially by a recruitment of mononuclear cells, neutrophils and eosinophils around the parasite egg; these are gradually replaced by histiocytes and epithelioid cells. The schistosome egg becomes degraded as a result of prolonged immune cell activity and collagen fibres become more prominent. Finally, fibrocytes and collagen fibres become the predominant feature of the granuloma and it decreases in size. The egg disintegrates and may become calcified [35]. It is the progressive damage caused to tissue through granulomas that lead to periportal fibrosis, portal hypertension and the risk of haematemesis (vomiting blood). Hardening and enlargement of the liver or spleen, which are consequences of the chronic damage of granulomas around *S. mansoni* eggs, are estimated to affect 8.5 and 6.3 million people respectively in sub-Saharan Africa [36].

Global Distribution and Epidemiology

The epidemiology of *S. mansoni* is characterised by: i) stability of prevalence, both spatially and temporally, owing to the long duration of worms inside hosts; ii) a spatial distribution that is predictable at the macro-scale though heterogeneous at the micro-scale; iii) an overdispersed distribution of parasites within a community, in which a small number of hosts harbouring the majority of worms; and iv) an age- dependent distribution in hosts.

Stable endemicity

Infectious diseases can be considered to fall on a continuum between those from which cases predominantly arise during epidemics or outbreaks and those that are predominantly endemic in the human population. The former often arise from pathogens that are harboured in environmental or animal reservoirs before ‘spilling over’ and infecting humans, and may spread from person to person until the outbreak ends or is contained; Ebola and Lassa virus are examples [37]. Endemic disease, by contrast, is caused by pathogens that often have intermediate hosts, long

latency or durations of infectiousness such as Guinea worm, falciparum malaria and HIV. Epidemic diseases fluctuate in prevalence through time and space by orders of magnitude and their transmission is largely unpredictable [38, 39]. By contrast, the prevalence of endemic diseases within a population is essentially stable, therefore point prevalence surveys at any moment in time should provide a reliable estimate (assuming accurate diagnosis and random sampling). For this reason, endemic diseases such as soil transmitted helminths, lymphatic filariasis, malaria, and schistosomiasis have been subject to large-scale mapping projects that have been used to inform public health control programs [40-43]. In the case of schistosomiasis, this stability is provided by the long lifespan of the adult worms (on average 6-10 years [23]), greater than that of the snail (weeks) or free-swimming larval stages (hours), meaning that any fluctuations in seasonal transmission are minor in accounting for changes to prevalence and infection intensity [23, 44].

Spatial distribution

Covariates such as mean rainfall, altitude, poor sanitation, distance from water bodies and past treatment history with mass drug administration (MDA) explain much of the variance in the spatial distribution of schistosomiasis and can be used to estimate prevalence in areas where point prevalence survey data is absent [45]. Despite these spatially predictive variables, prevalence and infection intensity can be heterogeneous at small spatial scales; a study on the spatial variance of *S. mansoni* in East Africa found strong evidence for spatial clustering that was more pronounced than for hookworm infections in the same area [46].

Parasite aggregation within hosts

The infection intensity of adult *S. mansoni* worms is approximated by measuring parasite eggs in stool (see below), though the relationship between egg burden and adult worms is non-linear and can be affected by 'density dependence', as adult worm fecundity, for example, declines with the density of the infection within a host [47]. Within communities infected with *S. mansoni* the distribution of adult worms is highly variable among hosts and can be modelling through the negative binomial distribution, where the high variance can be accounted for with a dispersion parameter. This epidemiological pattern is also true for other helminths; as the distributions of parasites within hosts are highly aggregated the majority of worms are therefore harboured within a small number of "wormy people" [23, 48].

Age dependent infection intensity

Mean intensity of infection with *S. mansoni* in endemic areas shows an age-dependent profile, whereby infection intensity rises through the infant years to peak at around ten or eleven years of age and subsequently drops during adulthood. This finding has been replicated across study sites in Africa and the New World [49]. During the 1990s there was debate as to whether these age-dependent changes in infection intensity were due to a shifting immune response or differences in water contact time between children and adults [50]. The case was largely put to rest by studies in communities of fishermen where adults have a greater amount of water contact time than children, but which still showed the same fall in the number of eggs in stool following puberty [51].

Diagnostics

Parasitological measures were the first diagnostic measures to be employed in the diagnosis of schistosomiasis through detecting eggs in stool or urine and still remain the 'gold standard' for large-scale epidemiological surveys. The Kato-Katz thick smear developed in 1972 is a quantitative concentration technique widely recommended for the diagnosis of intestinal schistosomiasis and helminthiasis [52]. Stool is pressed through a sieve and a measured quantity transferred to a slide. A piece of cellophane soaked in glycerine-malachite green is pressed on the stool; the glycerine clears the faecal debris and the malachite green stains the eggs enabling examination of the slide 45 minutes or more after preparation. A key advantage of Kato-Katz is that, due to its extensive use in parasitology, technicians are confident in its practise. It also requires minimal equipment and can diagnose both intestinal schistosomiasis and soil-transmitted helminths (STH). However, the technique is messy and exposes the technician to an unnecessary risk of infection. The method is also affected by the weight and consistency of the stool collected. To increase the sensitivity of the method, stool samples may be used to produce duplicate or triplicate slides or repeated over numerous days, though this increases the cost of any diagnostic survey [53-55].

Other parasitological procedures have been proposed that involve formal-ether concentrate or salt flotation. These methods have been shown to improve sensitivity: for instance the FLOTAC method, which uses a zinc-sulphate flotation solution and a centrifuge to bring eggs to the surface

before reading under a light microscope, was found in field trials to have a greater sensitivity than triplicate Kato-Katz or formal-ether concentrate [56]. However the cost, equipment requirements and complexity of laboratory procedures prohibit the extension of these methods beyond small-scale studies into large-scale epidemiological surveys [57]. In the case of FLOTAC the use of a centrifuge (and hence electricity) means that samples cannot be read at the point of collection, but rather must be fixed in sodium acetate-acetic acid-formalin and transported to a larger laboratory, which is impractical when diagnosing thousands of patients in rural areas.

Immunological tests may target anti-schistosomal antibodies or schistosomal antigens in plasma, serum, urine or sputum. A general problem with immunological diagnostics is that they cannot readily discriminate between active infection, past infection and re-infection [57]. The urine-based circulating cathodic antigen (CCA) test for *S. mansoni* has proven useful for rapid baseline surveys of countries due to its high sensitivity and ease of use [58], however once treatment programs get underway antigen based tests currently do not have the specificity to determine whether an individual has been cleared of parasites or not [59]. The CCA test is only able to diagnose intestinal schistosomiasis, therefore separate point of care antigen diagnostic tests are required to detect *S. haematobium*. Another circulating antigen test (circulating anodic antigen; CAA) has also been developed as an ELISA and shows comparable sensitivity and specificity to CCA, though a rapid point of care test for CAA remains in development [60]. While there are concerns about the inadequacy of current diagnostics and the need for new or better tools [61], Kato-Katz remains the dominant method for determining the prevalence and infection intensity of communities, despite the limitations noted above, due to its simplicity and low cost. Moreover, it is the only method currently used to evaluate the effectiveness of anthelmintic treatment [62].

Treatment and Control

Interventions to control schistosomiasis have varied over time. Early national control programs, such as the Sudanese Blue Nile Health Project (started 1979) and the Egyptian program (started 1988) used a combination of chemotherapy, snail control, provision of clean water and sanitation and infrastructural development [63, 64]. However, these multi-approach control programs are now the exception rather than the rule [65]. Due to the low cost, safety and generally high efficacy of praziquantel, MDA has become the mainstay of national control programs.

The Schistosomiasis Control Initiative (SCI), founded in 2002 with funding from the Bill and Melinda Gates Foundation, was the first organization to implement a multinational MDA control program against schistosomiasis. The initial premises of the SCI were that delivery of MDA to infected children “two or three times during their school life” would prevent the severe morbidity caused by conditions such as ascites (the accumulation of fluid in the abdominal cavity) and hepatosplenomegaly [66]. In addition deworming would improve the growth and nutritional status of children [67]. Uganda was chosen as the first country where the SCI would operate, due to political will within the country and the severe morbidity known to exist in certain regions; hospital records from the West Nile District from 1951 to 1978 showed schistosomiasis to be the fifth most common cause of admission and a major cause of death [68]. Following the start of MDA, longitudinal studies examining the impact of treatment with praziquantel in Uganda on 1871 children from 37 schools found that after two years *S. mansoni* infection intensity had dropped by 70% and the children had shown significant improvement across a range of health outcomes such as haemoglobin concentration and a reduction of clinical morbidity in the liver and spleen [69].

In the strategic plan for 2012-2020, the World Health Organisation (WHO) outlined its intention to control morbidity due to schistosomiasis by 2020, eliminate schistosomiasis as a public health problem by 2025 in all endemic countries and interrupt transmission in all regions outside of Africa and certain countries within Africa by 2025 [70]. This ambitious plan is solely reliant on the use of praziquantel in MDA programs. While reduction of transmission was not one of the initial goals of SCI, there is evidence from Uganda that MDA with praziquantel alone can reduce transmission. From 2003-6 the ‘force of infection’, a transmission parameter equivalent to the *per capita* rate of infection, was significantly reduced after three rounds of praziquantel when compared to baseline as inferred from egg counts before and after three years of annual MDA [71]. A sustained reduction to the prevalence and intensity of infection with *S. mansoni* and *S. haematobium* has also been observed in Burkina Faso following a decade of MDA. In the 1980s the prevalence of *S. mansoni* in western districts of Burkina Faso was as high as 79%; following an MDA program started in 2004, the prevalence in these areas had dropped to $\leq 1\%$ by 2013 [72, 73]. The prevalence of *S. haematobium* has also significantly declined in most districts, suggesting that schistosomiasis can be controlled in this country through MDA with praziquantel alone.

Praziquantel

Praziquantel is an anthelmintic based on the anti-parasitic pyrazino isoquinoline ring system. The drug was originally synthesised in the laboratories of E. Merck in Darmstadt, Germany in the 1970s [74]. It is a white crystalline powder consisting of an equal mix of optical isomers, of which only the 'laevo' (anticlockwise) form has schistosomicidal activity [75]. Orally administered praziquantel is rapidly absorbed into the blood and peak levels occur after 1-2 hours in normal volunteers. The toxicity of praziquantel is low, however side-effects are reported in field treatment studies [76]. Generally side-effects are correlated with the intensity of infection and are likely to be an immune response to products released from the death of adult worms. Side-effects are usually minor, such as headaches, nausea, vomiting and abdominal pain and are transitory with most disappearing after 1-2 hours [76, 77].

Despite many years of study, praziquantel's precise mode of action remains unknown. The immediate effect of the drug on schistosomes is spastic paralysis of worm musculature, which is accompanied, and possibly caused by, a rapid Ca^{2+} influx into the schistosome muscle cells [78]. When the action of Ca^{2+} channels was blocked by two inhibiting compounds (nicardipine and nifedipine), *S. mansoni* was found to have a 50% greater chance of survival in a concentration of praziquantel that is usually fatal [79]. Furthermore, there is an immune dependent component to the action of praziquantel. Brindley and Sher found that B-cell depleted mice were only 20% as effective in clearing *S. mansoni* after administering praziquantel compared with normal mice [80].

Drug resistance in praziquantel?

The first study to document the potential for *S. mansoni* to develop resistance to praziquantel was carried out by Fallon and Doenhoff, who exposed *S. mansoni* infected mice to increasing concentrations of praziquantel and oxamniquine; the latter is an *S. mansoni*-specific drug used predominantly in Brazil. The authors' found that after 6 and 7 rounds of exposure to praziquantel 80% and 93% of adult worms survived three doses of 300mg/kg of praziquantel, and after 6 rounds of exposure to oxamniquine the *S. mansoni* worms were completely tolerant. It was also found that resistance to either praziquantel or oxamniquine did not confer resistance to the other molecule, supporting separate modes of actions for both compounds [81].

In the field, drug efficacy cannot be observed directly as the adult worms are within the host. Therefore, efficacy must be estimated from diagnostic tests before and after treatment, usually, as discussed above, in the form of egg counts. Sample estimates are typically then used to quantify drug efficacy; the cure rate (CR) and the egg reduction rate (ERR). Egg reduction rates are more informative than CRs as morbidity is linked to infection intensity and, in the case of schistosomiasis, praziquantel is not truly curative as the drug is unable to target juveniles. The World Health Organisation (WHO) has set a “tentative” ERR threshold of 90% as a target efficacy for studies investigating anthelmintic efficacy [82].

The use of sample estimates, whereby CRs and ERRs are calculated directly from the observed data using simple arithmetic operations, while being simple to calculate and easy to interpret, present a number of limitations [62]. Firstly, multiple readings of egg counts must be averaged out in order to obtain a single count per individual per time-point, this loses any information on the variance of egg counts or the differences in variation between individuals as patients may be present for different number of days during testing. Secondly, the approach is unable to formally test the importance of explanatory variables that may affect drug efficacy. Thirdly, sample estimates, as a ratio between two means, lack simple methods for calculating uncertainty around the mean [83]. Finally, sample statistics are unable to capture the effects of correlated observations (such as measures taken from a particular time point or within a school). A more powerful inferential approach makes use of mixed effects regression techniques that are suitable for analysing longitudinal data, though studies have only recently begun to implement these methods [84].

Despite the tendency of authors to speculate on the potential for drug resistance to develop during treatment programs [75, 85, 86], the evidence for reduced efficacy of praziquantel against schistosomes in field reports remains patchy. For instance, in Northern Senegal following the development of the Diama Dam in the 1980’s the prevalence of *S. haematobium* and *S. mansoni* increased [87] and reports emerged of extremely low cure rates in *S. mansoni*, suggesting the potential for resistance against praziquantel [76]. However, cure rates are strongly influenced by the starting intensity of infection, along with the sensitivity of diagnostics [88] and subsequent investigation in the Senegal River Basin found high ERRs of 97-99% [89]. Reduced efficacy of

praziquantel against *S. mansoni* has also been reported in Egypt [90] with ERR's as low as 71%, though it is unclear whether these represent high rates of reinfection or anthelmintic resistance [91]. This thesis presents a contribution to understanding the reduced efficacy of praziquantel against *S. mansoni* in Uganda, which forms the basis of Chapter 3 and the publication is presented in Appendix 4 [92].

Past population genetic studies of *S. mansoni*

Darwin established that natural selection acts upon traits that are variable and heritable within populations [93]. While variation within a species can be observed phenotypically this is underpinned by genetic or molecular variation that can take myriad forms within a genome, from large structural variations spanning kilobases (kb) or rearranging chromosomes, down to single nucleotide base changes or polymorphisms (SNPs).

The extent to which the genetic variability in populations of organisms can be observed has been restricted by the technologies available to researchers. Initially, genetic variation was characterised by electrophoresis studies of proteins or DNA [94], then with the advent of Sanger sequencing and the polymerase chain reaction (PCR) this improved genetic analysis to the resolution of several hundred base-pairs (bp), typically applied to highly variable regions of the genome [20, 95]. Similarities between these sequences were then used to infer relationships between individuals within a population or between populations along with events in the history of a species such as speciation or changes to effective population size such as through a bottleneck.

In 2003 Paul Hebert and colleagues published a paper stating that the Cytochrome Oxidase 1 (*cox1*) mitochondrial gene could be used as a universal barcode for life, a “silver bullet” to aid taxonomists in species recognition across the animal kingdom [96]. This locus was chosen due to its relatively high rates of molecular evolution, the availability of robust universal primers and a lack of insertions or deletions (indels). The stated aim of the barcoding project was that, in the event of an organism's species being unclear, it could be compared with the DNA barcode of a ‘type specimen’, which had been verified by taxonomists. It was not designed for the discovery of new species, to reorganise taxonomies, or for within-species analysis. Furthermore tree-building

analysis were discouraged this went beyond the technology's scope and has been described as among the “deadly sins” of DNA barcoding [97, 98].

Despite the stated aims of the barcoding project, the sequencing of *cox1* genes in schistosomes has been used widely for population genetics, starting with a ‘global population’ overview carried out by Morgan *et al.* in 2005 [20]. Concerns were voiced in 2007 that the *cox1* gene was unsuitable for population genetics in *S. mansoni* and fared poorly in recreating phylogenies inferred from full mitochondrial genomes. Zarowiecki and colleagues argued that “Considering the information content of entire mt genome sequences, partial *cox1* would not be the ideal marker for either species identification (barcoding) or population studies with *Schistosoma* species” [99]. Given these warnings, it is therefore surprising that numerous studies of within species variation of *S. mansoni* (and other schistosomes) have been undertaken in subsequent years on the basis of *cox1* gene (or partial fragment of the gene). While the problems of inferring population structure from a single gene are well known [100], this has not prevented Stothard and colleagues, for example, from speculating that the variability of *cox1* demonstrates a “significant population partition” between populations of *S. mansoni* from Lakes Victoria and Albert in Uganda [101], despite the phylogeny of this particular paper failing to show the two groups separating into distinct clades, nor from a study claiming that this locus is able to differentiate global *S. mansoni* into 5 clades despite weak bootstrap support [102].

Warnings on the use of single markers come from another genus of helminthes, *Ascaris* spp., where there has been confusion over whether *A. lumbricoides* and *A. suum*, which infect humans and pigs respectively, are in fact the same species [103]. Looking at two regions of ~300bp in the nuclear ribosomal DNA (rDNA) region, Zhu and colleagues inferred that the nucleotide differences between *A. lumbricoides* and *A. suum* were comparable to within species diversity and so questioned their separate taxonomic status [104]. Full sequencing of the mitochondrial genomes of both species revealed them to be highly similar, differing by only 1.9%, whereas the diversity found within *A. suum* was 1.5%, leading Liu *et al.* (2012) to concluded that the data “provide some support that *A. lumbricoide* [sic] and *A. suum* represent a single species” [105]. To help resolve the issue, Betson *et al.* (2014) sequenced both the *cox1* gene and a panel of 8 microsatellite markers from > 500 *Ascaris* spp. isolates from humans and pigs in 10 countries around the world, including the UK, Uganda and Guatemala [106]. The authors found that *cox1* was unable to differentiate

between human and pig populations, with certain 'haplotypes' being equally represented among worms from either host. However, a supervised clustering algorithm of microsatellite sequences showed that the data fell into 2 clusters that clearly separated human and pig *Ascaris* into separate populations. Further, the fixation index of microsatellite regions showed that worms from humans were more closely related across countries than were human and pig *Ascaris* from within the same country. This study demonstrates the danger of inferring taxonomic relationships from single markers compared with multiple markers, and that even the complete mitochondrial genome may not be representative of nuclear regions.

The multipurpose *cox1* gene has also been used in conjunction with a nuclear rRNA sequence to investigate hybridization between different species of schistosomes, including the human-infective *S. haematobium* and the ruminant-infective *S. bovis* and *S. curassoni* [107] and also between *S. mansoni* and the rodent-infective *S. rodhaini* [108]. While the capacity of schistosomes to hybridize has been known since the 1950s [109], for authors to claim on the basis of modest molecular evidence that "hybrids are not 1st generation and are a result of parental and/or hybrid backcrosses, indicating a stable hybrid zone" [107] seems premature given the complexity of the patterns of hybridization that can be seen across whole genomes. For instance, the fungal species *Zymoseptoria pseudotritici* was found to be the product of hybridization between two other species that resulted in a mosaic of 100 kb regions along the *Z. pseudotritici* genome that alternate between being highly variability and monomorphic, explained by a bottleneck proceeding the hybridization event [110]. Such a pattern could only be understood through whole genome sequencing and likewise higher resolution genetic and genomic data are likely to be required to infer the true nature of hybridization between schistosome species.

Genetic population structure of *S. mansoni* populations

Parasite populations are inherently structured into subpopulations (for instance, within hosts, known as 'intrapopulations') and the hosts may themselves show different levels of structure (ecological areas, villages, schools etc). Therefore, the impact of this population structure on parasite population genetics is an important consideration. A study by Criscione and Blouin (2006) on different species of trematodes that infect salmon aimed to compare species with an autogenic lifecycle, where the parasites' hosts live exclusively in freshwater, with an allogenic

lifecycle, in which the trematode species has terrestrial hosts [111]. Based on results from a single mitochondrial gene, the allogenic trematode species *Nanophyetus salmincola* showed significantly higher rates of migration between subpopulations and only limited structuring compared with autogenic species (*Deropegus aspinia* and *Plagioporus shawi*), which showed lower rates of migration between subpopulations and greater population structuring. This suggests that parasite species' have greater rates of migration and gene flow where definitive hosts are able to disperse parasites between different subpopulations. This has implications for the population genetics of *S. mansoni*, which would be expected to show low levels of population genetic structure as it has a human definitive host. In a meta-analysis of microsatellite diversity in *S. mansoni* populations from five countries across East and West Africa, Gower *et al.* (2013) again found low levels of F_{ST} at different spatial scales ($F_{ST} \leq 0.113$), with the majority of variance in the population structure shown within schools or between miracidia within individuals, rather than over much larger spatial scales, such as between countries [95]. A separate study on adult *S. mansoni* which, unusually, have rats as their major definitive host in Guadeloupe also found low F_{ST} (≤ 0.07) on the basis of microsatellite markers [112].

Effect of praziquantel on *S. mansoni* populations

Along with “barcoding regions”, microsatellites have also been widely used in recent *Schistosoma* evolutionary analysis, including population genetic studies examining potential changes to microsatellite diversity of *S. mansoni* before and after treatment with praziquantel. A study in Tanzania from 2005-6 [113] based on 7 microsatellite markers sampled 80 praziquantel naïve children before and 47 children one year following treatment (total miracidia = 1706). The results revealed a significant reduction in the allelic richness and expected heterozygosity of *S. mansoni* following treatment within individuals. Further, the authors found the same reduction in diversity among untreated individuals at follow up. This suggested that after a single round of treatment with praziquantel some degree of selection may be acting on the parasites and that only a subset of the initial *S. mansoni* population were able to survive treatment or reinvade. The authors noted “this result tends to indicate that not only did one round of MDA affect parasite diversity within each child treated and across all children treated; it may also have affected parasite diversity within untreated children and therefore the broader *S. mansoni* population as a whole” [113].

However, these findings have failed to be replicated in a number of subsequent studies. In a study conducted in northern Senegal from 2007-8, Huyse *et al.* (2013) analysed microsatellite markers from miracidia collected from 12 children at baseline, 6 months follow up (after 2 rounds of MDA) and 2 years follow up (after 5 rounds of MDA) [114]. Over this period there were no significant changes to allelic richness or expected heterozygosity in *S. mansoni*. This study may have lacked sufficient power to detect any changes to genetic diversity as modelling work has shown that increasing the number of hosts, rather than the number of miracidia per host, improves the resolution to detect population changes before and after treatment [115]. Therefore, sampling from more than 12 hosts may be necessary to detect subtle changes to genetic structure. Some studies with greater statistical power have, nevertheless, also been unable to detect changes to *S. mansoni* variability following treatment. A four year cohort study in western Kenya that followed 67 individuals and genotyped a total of 4938 miracidia with 12 microsatellite loci also found no changes, or a slight increase, in allelic richness or expected heterozygosity each year following treatment [116].

It is important to consider that these studies are looking to demonstrate an effect of large-scale praziquantel treatment on the parasite population as a whole. In all cases follow up sampling was performed long after the initial treatment with praziquantel (≥ 1 year) and so such studies would be unlikely to detect differences between parasites that survive clearance and those that are sensitive to treatment. Even in the event that a fraction of *S. mansoni* parasites do resist or tolerate treatment with praziquantel, after a year or more of continued re-infections such genotypes may be indistinguishable against the background of population diversity. However, a microsatellite study in Brazil [117] where the similarity of *S. mansoni* populations that survived praziquantel only four weeks after treatment were compared with susceptible worms also found no significant difference between the groups, though a different measure of population structure, the “D index” [118], was used in the analysis, meaning the results are not necessarily comparable with other studies.

In addition, the absence of differentiation or change at neutral markers is not necessarily indicative of whole genome processes [119]. A study on Atlantic cod, for instance, showed that very little genetic differentiation was observed over nine microsatellite loci, but that substantial variation was taking place at the PanI locus, which was known to be under natural selection [120].

Next Generation Sequencing and Parasite Genomics

The field of genomics has been revolutionised by the advent of high-throughput sequencing, also known as next-generation sequencing (NGS), technology. After the sequencing of the human genome at enormous cost and over a period of ten years through Sanger sequencing, subsequent technological improvement, starting with the SOLiD technology in the early 2000s, decreased the cost of sequencing by a factor of 50,000 [7].

While a number of competing technologies exist, the dominant sequencing technology at present is manufactured by Illumina [121]. If one were to imagine an idealised sequencing process this would involve: i) long reads that spanned the largest repeats; ii) retention of the haplotypes of variants; and iii) a low error rate. If we compare the current Illumina technology against our expectations it is characterised by sequencing extremely short reads (70-150bp), an absence of haplotype information due to the short read lengths (hence variants often require statistical phasing) and a relatively high error rate ($\sim 0.1\%$). While emerging technologies (so called “third generation sequencing”) hold promise in overcoming some of these shortcomings, such as long read technologies pioneered by Pacific Biosciences and Oxford Nanopore, due to affordability and a wide-range of platforms, the Illumina ‘suite’ of instruments remain the dominant mode of sequencing in most labs worldwide and subsequently a great deal of bioinformatics research focuses on overcoming the problems inherent in this technology.

The process of Illumina sequencing can be briefly outlined as follows. First, purified DNA is sheared into fragments of a known length, typically several hundred base pairs. This can be achieved through sonication and the DNA fragment sizes (or ‘insert sizes’) form a normal distribution around the desired length. Adapters are then ligated onto the ends of the DNA fragments, and depending on the concentration of the DNA, a number of PCR cycles then follow to increase the amount of DNA prior to sequencing. DNA is then loaded into the sequencing machines on glass flow cells. Adapter-bound DNA fragments are immobilised on the flow cell surface, which contains oligonucleotides that bind to the adaptors on the DNA template. Then a process known as ‘solid phase amplification’ leads to the initial DNA template bridging onto adjacent oligonucleotides and a polymerase binding to the template and building a complementary strand by incorporating dNTPs in solution. The double stranded DNA is denatured and so the forward and reverse strands are left attached separately to oligonucleotides anchored to the nano-well.

This process is repeated so that around 1000 identical DNA fragments are formed within a single well, this step ends with the reverse strands being cleaved and washed off, leaving only the forward strands. Across the flow cell this process takes place within each well simultaneously so that billions of wells with monoclonal DNA clusters are formed.

The sequencing that follows is known as 'sequencing by synthesis'. Nucleotides that are tagged with fluorophores bind to the DNA fragments within each cluster. As the correct base is incorporated a characteristic fluorescence is emitted and lasers capture the signal. Within each cluster each clonal DNA template is read simultaneously. The sequencing continues in a number of pre-defined cycles corresponding to the read-length. Information on every paired read is typically output as a 'fastq' file, though the file is unordered. To make sense of the reads, they must be assembled *de novo* or aligned to a reference genome [121, 122]. Due to the relatively high error rate and short read length of NGS technology, to reliably call a SNP or indel in sequence data requires any variable position in the genome to be covered by multiple independent reads. If many independent reads suggest that a variant is present in the sample compared to the reference genome then it can be called reliably [123]. For this reason, it is important to have high coverage (also referred to as depth) when calling variants across a population.

The *S. mansoni* genome consists of seven autosomal chromosomes and a ZW sex determination system, in which males are homogametic with ZZ sex chromosomes and females are heterogametic with ZW sex chromosomes. The W chromosome is smaller than the Z chromosome, though to-date the genome assembly of *S. mansoni* has yet to fully resolve this region as the W chromosome is highly repetitive [124]. A draft 380 megabase (mb) genome of *S. mansoni* was first presented in 2009 [125]. The genome was found to contain unusual features such a large proportion (20%) of long terminal repeats and other transposable elements along with 45 unusually small micro-exon genes (MEGs) with exons of 6-36 basepairs flanked by conventional exons at the 5' and 3' end of the gene. While the MEGs did not appear to have a common function, it was proposed that their structure allowed for splice variants due to exon skipping, thus increasing the possible protein variation. Micro-exon genes were also found in the genomes of four tapeworm species, suggesting the feature is common to Platyhelminthes [126]. The reference genome was systematically improved by the addition of high coverage Illumina reads and capillary sequencing along with manual inspection, leading to a second publication in 2012 which

described a draft assembly decreasing in size from 374.9 to 364.5 mb, with 81% of the content assigned to chromosomes (increased from 55%) and the total number of contiguous sequences (contigs) decreased from over 50,000 to 9,200 [124]. Improvements to gene models were also made through sequencing the transcriptome at different life stages, this resulted in the number of predicted genes dropping from 11,807 to 10,852. Draft genomes of *S. japonicum* and *S. haematobium* were published in 2009 and 2012 respectively, though no subsequent efforts to improve the assemblies for these species have yet been published [127, 128].

Despite the availability of these genomic resources for schistosomes, there has been only modest progress in utilizing them for population studies. In one of the first examples using NGS to explore larger genomic regions, seven pools of *S. japonicum* from China were sequenced. Phylogenetic reconstruction using genes shared between *S. japonicum* and *S. mansoni* found that the population structure reflected geographic proximity, and was broadly similar to that found in the original 1999 electrophoretic study from which the samples originated [129, 130]. The authors performed an enrichment study on the most variable genes, though poor functional annotation limited the extent to which they could infer biological meaning from these genes. Another research group has developed a method for capturing the 15 mb exome (exons of genes) of *S. mansoni* for NGS sequencing [131], though their work applying this method to the population structure of *S. mansoni* is yet to be published. This thesis presents the first population genomic study of *S. mansoni*, which is the subject of Chapter 2 and the publication is presented in Appendix 3 [132].

Exploring drug resistance with genomics

One of the most important prospective uses for parasite genomic resources is the potential to explore resistance to drugs. In a recent example that harnessed the power of quantitative genetics, parental *S. mansoni* parasites that differed 500 fold in their susceptibility to the drug oxamniquine were ‘intercrossed’ to produce F1 and F2 progeny [133]. The proportion of susceptible F2 progeny (~25%) confirmed a recessive Mendelian inheritance of the trait; 144 F2s were genotyped with 62 microsatellite markers distributed in 20 centimorgan (cM) intervals across the genome, which subsequently identified a region at the end of chromosome 6 as having a high LOD (logarithm of odds) score. Whole genome sequencing of the parent worms and F1 progeny identified >500,000 SNPs, including >6,000 SNPs in the region of interest. Of these variants 48 SNPs were chosen along

the genome, including 17 in the region of interest to improve the resolution; this increased the LOD score and narrowed the QTL support region to 439 kb. The 16 genes within this region were then narrowed down using additional evidence, such as RNA-seq expression data and the presence of a non-synonymous SNP, to a candidate list of 6 genes. Finally, a sulfotransferase gene was implicated through biochemical assays and RNAi knockdown.

The possibility of extending this approach to mapping genes conferring resistance to other anti-parasitic drugs, however, is restricted by the nature of the trait. Laboratory induced resistance to praziquantel for instance, seems to confer a 5-fold difference in susceptibility to the drug [86], much lower than the 500-fold difference induced in oxamniquine resistant worms, and so the power to detect such a trait in an experimental cross would be reduced, hence requiring more progeny or more markers to be sequenced. Furthermore, a single Mendelian gene confers oxamniquine resistance meaning that it was more straightforward to infer the genetic architecture than in the case of a complex genetic trait that is underpinned by numerous SNPs.

The evolution of resistance genes under field conditions has been the subject of theoretical consideration, though until recently little data has been available to study the epidemiological and genomic changes to a natural population under drug selection. With recent developments in genomics along with enhanced epidemiological surveillance, we can now compare real world scenarios against our expectations. A useful case study is provided by the protozoan parasite *Plasmodium falciparum*; as the development of resistance to artemisinin in South East Asia has been of great concern and prompted genomic scans to determine the genetic basis for resistance.

If we consider a situation whereby a susceptible allele S and a resistant allele R exist at frequencies p_t and q_t respectively at generation t , then $p_t + q_t = 1$. Genotype RR is taken to be resistant and SS to be susceptible and heterozygotes RS of intermediate fitness. In the presence of drug pressure, the fitness of the three genotypes are denoted by w_{RR} , w_{RS} , w_{SS} , whereby $w_{RR} > w_{RS} > w_{SS}$. Gene frequencies in a successive generation are given by:

$$p_{t+1} = \frac{w_{RR}p_t^2 + w_{RS}p_tq_t}{w_{RR}p_t^2 + 2w_{RS}p_tq_t + w_{SS}q_t^2} \quad (1)$$

The initial proportion of q_t is assumed to be low in the population, though tends to fixation as t increases to infinity in the absence of heterozygote advantage; plotting q against time results in a sigmoidal S-shaped curve. The implications of this are that the resistance phenotype may only be detected after many years of control during which the resistant allele is at a low frequency [23]. While the equation pertains to a diploid organism, *P. falciparum* can be considered diploid at the level of individual human hosts as multiple, genetically variable infections exist within a single patient and so genomic sequencing is able to capture within-host parasite diversity [134].

These theoretical predications have been borne out by longitudinal data on *P. falciparum* clearance half-lives in north-western Thailand. Over the period from 2001 to 2010 parasite clearance times rose from a mean of 2.6 hours in 2001 to 3.7 hours in 2010, and the proportion of slow clearing infections increased from 0.6% to 20% over the same period [135]. The alarm over artemisinin resistance in Thailand was only raised in 2009 [136], though the retrospective analysis showed that the ‘creep’ of resistant isolates was evident before this time [135].

Once the phenotypic nature of artemisinin resistance had been established, the race was then on to determine its genetic basis. Cheeseman *et al.* (2012) exploited differences in parasite clearance rate between populations to infer the extent to which selective sweeps had acted in these populations [137]. As polymorphisms under selection increase in frequency, the long-range haplotype, that is the contiguous portion of a chromosome inherited during meiosis, will not have time to be broken up through recombination. Originally developed to compare human populations under the HapMap project, the cross population extended haplotype homozygosity (XP-EHH) test uses differences between populations to detect alleles that have reached fixation in the population under selection, but remain polymorphic in other populations [138]. By 2012, resistance to artemisinin had been established in western Cambodia and had recently emerged in Thailand. However, artemisinin resistance was absent from other Asian foci such as Laos. Ninety-one parasites were genotyped with 6,969 SNP markers from Cambodia, Thailand and Laos and XP-EHH analysis identified thirty-three SNPs under selection covering 2.4% of the genome. These regions were subsequently fine-mapped with an increased density of SNP and microsatellite markers in parasites from Thailand of which 4 markers spanning 35 kb showed a strong association with clearance rate. The causative mutations, SNPs within the C-terminal K13-propeller domain, were later detected in this region through an experiment in which the genomes

of *P. falciparum* isolates that had been exposed to gradually increasing doses of artemisinin over 125 cycles of drug pressure were compared with a control strain [139].

This example demonstrates how genomic detective work can isolate causative variants underpinning phenotypic parasite traits. Through whole-genome sequencing of schistosomes linked with epidemiological datasets, it may be able to uncover the genetic basis of important traits of *S. mansoni* [140]. However, unlike in genetic crosses where the background of the parasite population is controlled within the experiment, association mapping in natural populations must contend with background variation and population structure that has the potential to confound results. Therefore enough statistical power must be generated to find regions of the genome with signatures of selection above the noise of background variability [141].

Thesis outline

The aim of this thesis is to combine epidemiological and genomic datasets on *Schistosoma mansoni*, therefore the chapters that follow from this introduction detail a population genomic study of *S. mansoni* using global isolates (Chapter 2), an epidemiological study of *S. mansoni* from data collected in Uganda on egg reduction rates following treatment with praziquantel (Chapter 3) and finally a study combining epidemiological and genomic datasets by whole-genome sequencing miracidia from patients in Uganda and using epidemiological metadata to understand the influence of MDA on the population structure of *S. mansoni* and locate regions of the genome under selection that are associated with praziquantel response (Chapter 4).

Chapter 2 presents the first whole-genome sequence data analysis of *S. mansoni*, using 10 global isolates and an isolate of *Schistosoma rodhaini* was also included in the analysis for comparative purposes. I developed the pipelines for taking raw reads and mapping to the reference followed by SNP calling along the genome and filtering. The SNP data were applied to construct a phylogeny and principal component analysis to infer the population structure. To understand the evolutionary history of the populations, I applied coalescence models to whole-genome data to infer the dates of: i) the time to most recent common ancestor (TMRCA) between *S. mansoni* and *S. rodhaini*; ii) the spread of *S. mansoni* across sub-Saharan Africa; and iii) the migration to the New

World. Finally, I documented signatures of directional selection between *S. mansoni* and *S. rodhaini* to uncover genes involved in host adaptation and also under balancing selection in *S. mansoni* through the use of population genetic tests that compare within species polymorphism to between species polymorphism.

Chapter 3 presents epidemiological data from fieldwork I conducted in primary schools in eastern Uganda that had been subject to different levels of drug pressure from MDA with praziquantel. Children were examined for *S. mansoni* with parasitological diagnostic tests up to 3 days before and after treatment with praziquantel. I fitted a mixed-effects model to the egg count data to account for both fixed effects (MDA treatment intensity, sex, weight, co-infection with STH) and random effects (individual host, school). After marginalising over the coefficients, individuals from schools that were treated most intensively were found to have a lower estimated mean egg reduction rate compared with individuals from less intensively treated schools, suggesting that more rounds of MDA may be resulting in *S. mansoni* parasites that are better tolerating anthelmintic treatment.

Chapter 4 presents the results of collecting miracidia from patients in Uganda and developing a method for extracting and amplifying *S. mansoni* DNA from Whatman FTA® cards prior to sequencing. Whole-genome sequence data was generated for 213 miracidia from 35 patients. The paired-end reads were aligned to the V6 *S. mansoni* reference genome and SNPs called across the cohort prior to filtering. Using a dataset of 5.4 million SNPs, I documented the population structure, characterised regions of the genome under selection and associated SNPs against the clearance phenotype. I uncovered the strongest signal of selection and the highest association with the host clearance phenotype in a 256 kb region at the beginning of chromosome 2 that contains two calcium channels, suggesting that this region may contain variants underlying the observed differences in the anthelmintic efficacy of praziquantel.

Finally, the general discussion of Chapter 5 aims to place my findings and results from all chapters in a broader context and suggest areas for future work. Published papers from my thesis to date are provided in Appendixes 2-4.

Chapter 2

Whole genome resequencing of global *Schistosoma mansoni* isolates reveals population history and evidence of selection since divergence with *S. rodhaini*

Abstract

This chapter presents the first application of population genomics to *Schistosoma mansoni* based on high-coverage resequencing data from ten global isolates and an isolate of the closely-related species *Schistosoma rodhaini*, which infects rodents. Since the sequencing of the *S. mansoni* reference genome there has been an absence of population genomic studies investigating selection and the population history of the species at the resolution of genome-wide variation.

Isolates of *S. mansoni* were selected based on existing phylogenetic data to represent the extant diversity of the species. The ten samples were sequenced to a median depth of 46x and mapped to the *S. mansoni* v5.2 reference genome before calling between 0.5 and 2.0 million high-confidence SNPs for each isolate. A principal components analysis shows the isolates from Uganda (Lake Victoria and Lake Albert) to be the most differentiated and a maximum-likelihood phylogeny indicated that these samples were the first to branch off from *S. rodhaini*, supporting the theory that *S. mansoni* originated in this region. Using the McDonald-Kreitman and Hudson-Kreitman-Aguade tests, I documented genes under directional and balancing selection in *S. mansoni* since the divergence with *S. rodhaini* that may have facilitated adaptation to the human host. Genes involved in human invasion (cercarial elastase) were found among the 13 (of 9514) coding sequences that showed significant evidence of positive selection. Coalescence modeling revealed the speciation of *S. mansoni* and *S. rodhaini* as 107.5-147.6KYA, a period which overlaps with the earliest archaeological evidence for fishing in Africa. The effective population size of *S. mansoni* was estimated to decline between 20-90KYA, at a time when climatic changes caused lakes in East Africa to dry up, before rising as *S. mansoni* dispersed across the African continent during the Holocene. In addition, evidence from coalescence models found strong evidence to support the hypothesis that *S. mansoni* migrated to the New World with the 16-19th Century Atlantic Slave Trade.

This Chapter therefore presents a detailed understanding of the evolutionary history of *S. mansoni* since its emergence approximately 120KYA and how the parasite has been driven by climatic and human factors to adapt successfully to a changing world.

Introduction

Since the sequencing of the 380 megabase (mb) *Schistosoma mansoni* reference genome in 2009 [125], genomic resources have contributed to comparative [126] and functional studies [142], including the discovery of the molecular basis for oxamniquine resistance [133], but little is known about genome-wide variation between parasites in the field. While large populations of *S. mansoni* have been investigated [101, 102, 113], all previous studies in this field have been restricted to analysing a small number of markers, typically using the mitochondrial *cox1* gene, nuclear ITS region or <10 microsatellite loci. In this study I present the first population genomic study of *S. mansoni* based on whole genome re-sequencing data for nine isolates collected from field sites across Africa and the New World. African individuals were selected from distinct clades based on previous phylogenetic data to represent the diversity of extant *S. mansoni* [102]. The NMRI (Puerto Rican) laboratory strain that was used to build the reference genome is also included in the analysis [125] along with the whole-genome sequence of *Schistosoma rodhaini*; a rodent schistosome species that is firmly established as the closest outgroup to *S. mansoni* [143, 144].

Whole-genome resequencing of populations allows us to explore how patterns of genetic variation change across the genome and thus detect the effects of selection. This can be of particular importance in pathogens where genomics can identify targets of balancing selection that interact with host immune-systems, for instance genes involved in host cell invasion in *Plasmodium* spp. [145-147]. Testing for coding sequences under strong positive selection may identify genes involved in longer-term selection in a coevolving host-parasite system, for instance mutations under selection on toll-like receptor proteins in rodents provides protection from bacteria such as *Yersinia pestis* [148]. To shed light on the divergence between *S. mansoni* and *S. rodhaini*, I catalogued the extensive genome-wide variation between a global collection of isolates and analysed this variation to better understand the emergence of the human-infective *S. mansoni* and its subsequent demographic and evolutionary history. Using a set of population genetic approaches, I looked at how signatures of natural selection vary between different coding regions

of these genomes, aiming to identify genes that may be involved in host-pathogen interaction or may have facilitated the adaptation of *S. mansoni* to human hosts.

Understanding the evolutionary origins of pathogens such as *S. mansoni* and their subsequent demographic history has implications, not only for a better understanding of the parasite itself, but also of human evolutionary history as infectious diseases have consistently been shown as a major force driving selection in human populations [138, 149]. Despite its importance, little is known about either the origin or subsequent history of *S. mansoni*. Previous work based on interspecies divergence at two short loci estimated that the split between *S. mansoni* and *S. rodhaini* occurred somewhere between 1.2 and 10 million years ago (MYA), around the time that the hominin clade split from other great apes and that the human infective species emerged in West Africa [150]. This has been contested by the view that East Africa is the most likely origin of *S. mansoni* due to the “conspicuous amount of diversity” observed in sequence data from this region [20, 95, 102]. A subsequent study using a mitochondrial molecular clock suggested that the most recent common ancestor of extant East African *S. mansoni* occurred 0.3-0.43MYA and that the divergence between *S. mansoni* and *S. rodhaini* was around 2.8MYA [20].

Two main hypotheses have been proposed to explain the divergence between African and South American *S. mansoni*. One suggestion is that New and Old World *S. mansoni* diverged along with their intermediate snail hosts *Biomphalaria pfeifferi* and *B. glabrata*, the split between which is approximately dated to the separation of the continents 80-120 millions years ago [151, 152]. A second proposal is that the divergence between these two *S. mansoni* populations is much more recent as African populations entered the New World through the Atlantic Slave Trade of 16th-19th Centuries [102]. This hypothesis has been supported by studies using DNA barcoding and microsatellite techniques, which have observed less genetic diversity among West African and New World *S. mansoni* suggestive of a more recent introduction of the parasite into these areas [20, 95].

Coalescence theory seeks to explain population genetic parameters among samples such as time to most recent common ancestor (TMRCA) and effective population sizes (N_e) in order to infer demographic history from sequence data [153]. In contrast to data from a small set of loci, whole genome sequence data offers a more powerful and unbiased approach to reconstruct the history

of a species as recombination gradually breaks up the genome into blocks that have independent ancestry. Genome-wide data thus provides many independent estimates of the coalescent process of ancestry and far greater precision in estimates of population genetic parameters [154].

Two different coalescent models were employed in this chapter that both make simplifications in order to calculate the coalescence under recombination from whole-genome sequences [155]. Pairwise Sequentially Markovian Coalescent (PSMC) uses the pattern of heterozygosity along a single diploid genome to estimate N_e through time [156] and Generalised Phylogenetic Coalescent Samples (G-PhoCS) uses data from 1kb non-recombining windows across the genome so that the history of population divisions and migration between populations can be modeled [157]. These complementary methods have already been shown to infer the complex demographic history of species such as humans and great apes using a small number of individual genomes [158], though this is the first study to apply these methods to sequence data from a human-infective parasite.

Methods & Materials

Samples

Samples (Table 1), with the exception of NMRI, came from the schistosomiasis collection (SCAN) at the Natural History Museum (NHM), London [159]. All schistosomes provided by SCAN were collected by NHM staff in collaboration with in-country partners, except for the Guadeloupe specimens, which were originally provided by Dr André Thèron, Université de Perpignan, France. SCAN isolates were collected from field sites within infected *Biomphalaria* snails, which were shed in the laboratory and the cercariae used to infect mice. SCAN and Guadeloupe isolates were passaged between 2 and 15 times. The NMRI sample is a Puerto Rican isolate from the early 1940s that has subsequently been maintained in the laboratory for hundreds of generations[94] and was used to build the reference genome for *S. mansoni*. Adult worms were extracted from mice and kept in liquid nitrogen storage prior to DNA extraction. For each isolate a single adult worm was used for DNA extraction and sequencing.

Sequencing

The libraries were sequenced on the Illumina Genome Analyser IIX for 76 cycles in each direction (paired-end reads) using V4 SBS sequencing kit and proprietary reagents according to manufacturer's recommended protocol (<https://icom.illumina.com/>).

Variant Calling Pipeline

Reads were mapped against v5.2 of the *S. mansoni* reference genome [124] using smalt (<http://www.sanger.ac.uk/resources/software/smalt/>) using options (-a -o "-i 2000 -r 1 -x -y 0.85" -k 13 -s 2). This resulted in reads mapping to the region of highest similarity in the reference genome above an 85% similarity threshold. When reads could be mapped to multiple locations with equal similarity one position was randomly chosen, though these repetitively mapped reads are excluded by variant calling. Processing of the resulting bam files, prior to variant calling, was performed using GATK version 2.6 to remove duplicate reads and perform local realignment around indels [160]. Single nucleotide polymorphisms (SNPs) were called using three algorithms; HaplotypeCaller and UnifiedGenotyper (implemented in GATK) and samtools v0.1.19 mpileup [161]. Filtering was applied to each set of variants; SNPs with genotype quality (GQ) scores (a measure of the confidence of the SNP call) < 10 and read depth (DP) < 10 were removed. Lastly, the intersection of the three calling algorithms was found using bedtools version 2.17.0.

d_N/d_S ratio

The 11,727 coding sequences in the *S. mansoni* genome were filtered by removing alternative transcripts (1005 sequences removed), genes with low coverage in the *S. rodhaini* alignment (<75% of bases covered at 10x, 719 sequences removed), genes with excessive coverage (>1.75 x mean depth, 475 sequences removed) and gene models with an incorrect number of nucleotides (not divisible by three, 14 sequences removed) leaving a filtered subset of 9514 coding sequences for downstream analysis. Maximum-likelihood estimates of d_N/d_S for each protein-coding gene were calculated by running codeml from the PAML package [162] in branch mode, where the phylogenetic tree for African *S. mansoni* and *S. rodhaini* was specified and partitioned into two clades, separating the branch leading to the *S. rodhaini* sequence (the 'foreground' branch) from all other branches between different *S. mansoni* samples. I ran two scenarios for each coding sequence, in which the value for d_N/d_S on the foreground branch was fixed at 1 (null hypothesis) or unfixed (alternative hypothesis) and the log likelihood (lnL) in both scenarios compared using the likelihood ratio test. The test statistic D is calculated as $D = -2 * (\ln L_{\text{fix}} - \ln L_{\text{unfixed}})$. The test

statistic is then compared against a Chi-Square distribution, where degrees of freedom (df) are equal to the difference in the number of parameters in the two models $df = \text{params}_{\text{unfixed}} - \text{params}_{\text{fixed}}$, for all coding sequences in the dataset $df = 1$. The significance threshold for the Chi-Square distribution was set at $p < 0.05$. Statistical overrepresentation analysis was conducted using PantherDB (v10.0) [163].

Hudson-Kreitman-Aguade (HKA) Ratio

The HKA ratio is the difference in the ratio of polymorphic (π) and fixed differences (K) within a gene. Where π/K is high, therefore, this indicates balancing selection (i.e. diversity within African *S. mansoni*) and when low and enrichment of interspecific nucleotide divergence between African *S. mansoni* and *S. rodhaini*. The HKA test was implemented through the use of previously developed software [164].

McDonald-Kreitman Test

Polymorphic differences within populations (African *S. mansoni*) and fixed substitutions between populations (African *S. mansoni* vs *S. rodhaini*) were identified using previously developed software [165]. The MK skew, or neutrality index, is defined as $\log_2 ((N_{\text{spoly}}/S_{\text{poly}})/(N_{\text{sfix}}/S_{\text{fix}}))$, where N s and S are the number of non-synonymous and synonymous sites within a coding sequence, and subscripts *fix* and *poly* refer to fixed differences between populations and polymorphisms within a population respectively. To avoid divide by zero errors, 1 was added to all scores. GO term enrichment was conducted in PantherDB [163] and enrichment of pfam domains was performed using a one sided Fisher's exact test in R v3.2.2.

Consensus fasta files

The phylogenetics software package RaxML and coalescence models PSMC and G-PhoCS require fasta sequences; thus a consensus sequence of the diploid genotypes was generated from the filtered vcf files using a modified perl script. Heterozygotes are represented by IUPAC ambiguity codes. These files were filtered to remove mitochondrial sequences, scaffolds not aligned to chromosomes and the scaffold representing shared sequence between the Z and W chromosomes [124]; this left the scaffolds aligned to the 7 autosomal chromosomes. To ensure loci included in G-PhoCS were unlinked and sufficiently small that little recombination occurs within loci, I examined the decay in linkage disequilibrium (LD) with distance between pairs of SNPs in the data using

SNP data in vcf files, which were phased using SHAPEIT2 (v 2.790)[166], and analysed using VCFtools (v0.1.11) --hap-r2 and found that LD decays to background levels for distances great than around 400 kb (Figure 7). To be conservative, the consensus calls were divided into 1 kb blocks separated by 500 kb. These same input regions were concatenated and used as input for RAxML.

Phylogenetic Tree

A Maximum likelihood (ML) tree was constructed using RAxML version 8.0.24 [167] under a GTR model of nucleotide substitution with a gamma distribution of rate variation across sites [167]. For each run, twenty randomly generated starting trees were used, and support for partitions are estimated as proportions of 100 bootstrap-replicates displaying the partition. The final ML tree with bootstrap values were visualised in FigTree v1.4.0 [168].

Principal Component Analysis

A matrix was constructed where every bi-allelic site in the genome is classified as homozygous reference (0), homozygous alternative (1) or heterozygote (0.5). This matrix was transposed in R version 3.1.3 and the PCA was performed with the prcomp function with all default settings retained.

Estimates of parameters for coalescent models

Mutation rate was calculated through a linear regression of the relationship between log-genome size and log-mutation rate [169]. The best fitting model for mutation rate (μ ; μ) in bp was $\mu = 2.585324 \times 10^{-10} \times g^{0.584}$ for genome size g in mb, giving an estimate of 8.1×10^{-9} per bp per generation (Figure 6). Generation time was estimated by adding the time taken for the parasite to start producing eggs after infection with human host (~42 days) [170] with the latency period in snails (~28 days) + time for transition between the two hosts (1 day) = ~71 days (0.2 years). There is a wide variation around the generation time estimates, as latency period in snails is temperature and species dependent within the *Biomphalaria* genus. To obtain effective population sizes (N_e) in numbers of individuals, I used the relationship $N_e = \theta / (4\mu)$ where θ is the uncalibrated population size. To obtain the divergence time in years (D), I used the relationship, $D = (0.2\tau) / \mu$ where τ is the uncalibrated divergence time.

Generalised phylogenetic coalescence sampler (G-PhoCS)

G-PhoCS is used to estimate several key population parameters; ancestral population size, divergence times and gene flow between species [157]. I ran a number of different G-PhoCS analyses to test how sensitive estimates of these parameters are to different assumptions about population structure and migration (Figure 8). Each simulation was run for a minimum of 10 million iterations, with at least 20 parallel chains and until convergence was reached (effective sample size (ESS) for all parameters >200). Samples from the MCMC chains were also manually inspected in Tracer (v1.5) to ensure chains had reached convergence.

PSMC

The Pairwise Sequentially Markovian Coalescent (PSMC) model uses the local density of heterozygous sites to infer the time to most recent common ancestor (TMRCA) for blocks of the genome, effective population sizes for discrete historical time windows and the positions of transitions between TMRCAs (i.e. boundaries of blocks) are also inferred from data. Modelling only a single diploid genome and using a computationally convenient approximation to the full coalescent makes treating the coalescent with recombination tractable for genome-scale data [156]. Because data from only a single diploid genome is used, the method loses accuracy more recently in time, eg for humans it is not informative about demographic events more recently than 20KYA. PSMC was used with data from each field isolate of *S. mansoni*, each time using parameters -N50 -t30 -r4 -p "26*2+4+6", with the distribution of TMRCA segments chosen to ensure at least 10 recombination events occur in each segment, as suggested in the PSMC manual, on a preliminary set of variant calls from mpileup for these data. Bootstrapping was performed 100 times for each sample (Figure 11).

Results

Genome-wide intra- and inter-species diversity

The genomes of 10 *Schistosoma mansoni* and one *S. rodhaini* isolate were sequenced to a median of 46x coverage. Sequence data were mapped against the *S. mansoni* reference genome (v5.2) [124]; coverage varied between 23x and 58x between samples (Table 1). The proportion of total reads mapped and of 'proper pairs' (reads mapped in the correct orientation and distance) did not vary significantly between *S. mansoni* isolates (from 75 - 97% for total mapping, 70 - 91% for proper

pairs) with the exception of isolate Guadeloupe 1 where fewer reads map against the reference (58% total mapped and 54% for proper pairs) due to contamination in the sample, whereby 24.8% of sequenced reads mapped against *Mus musculus*. Reads from *S. rodhaini* mapped well against the *S. mansoni* reference suggesting that SNPs between species did not affect placement of reads, but the proportion of proper pairs was lower, presumably due to structural differences altering the orientation and distance between read pairs between the two species (84% of *S. rodhaini* reads map and 56% as proper pairs). These results confirm that the reference assembly is suitable for studying variation between all isolates. Additional evidence that the *S. mansoni* assembly captures most of the *S. rodhaini* gene complement comes from examining the depth of reads mapping to coding sequences: of 10,722 coding sequences in *S. mansoni*, 10,003 were covered by a minimum of 10x across at least 75% of the coding sequence in the *S. rodhaini* alignment.

The numbers of SNPs for each sample, based on the intersection of three SNP calling and filtering algorithms (see Table 1; Methods), show a general geographical gradient, with the number of SNPs increasing with distance from Puerto Rico, which is the origin of the isolate used to construct the reference genome (Figure 1). The total number of SNPs per sample (including both homozygous and heterozygous sites) ranged from $<1 \times 10^6$ across the New World and Senegal sequences, to $\sim 1.5 \times 10^6$ SNPs in Cameroon and $\sim 2 \times 10^6$ SNPs in Uganda. The outgroup species *S. rodhaini* has $\sim 7 \times 10^6$ fixed differences when called against the reference, representing an average of about one variant every 50 bp. A notable exception to this geographical trend was the Coastal Kenya sample, which showed a reduced number of SNPs ($\sim 8 \times 10^5$), more comparable to West African and New World strains than the more divergent Uganda sequences.

ID	Country	Year	Reads (x10 ⁹)	Median Coverage	MP	UG	HC	Overlap	Heterozygotes (%)
NMRI	Puerto Rico	1940	16.68	43	547791	585813	532965	459938	265341 (57.7)
Senegal	Senegal	1998	22.44	57	807680	843282	764845	661528	118616 (17.9)
Cameroon	Cameroon	1997	16.51	42	1555732	1599021	1491934	1332792	879648 (66.0)
Coastal Kenya	Kenya	1996	22.97	58	931195	971723	891959	793274	373547 (47.1)
Lake Albert	Uganda	2004	21.76	55	2236138	2292720	2142399	1951022	1216415 (62.3)
Lake Victoria	Uganda	2005	20.09	50	2384792	2438722	2281572	2043541	1241072 (60.7)
Guadeloupe 1	Guadeloupe	1983	8.20	23	966028	987299	933593	772668	508996 (65.9)
Guadeloupe 2	Guadeloupe	1983	10.90	30	983954	1012647	951671	830275	541482 (65.2)
Guadeloupe 3	Guadeloupe	1983	12.62	35	1046797	1083754	1009148	872007	557070 (63.9)
Guadeloupe 4	Guadeloupe	1983	8.56	24	961926	986457	927716	750295	470137 (62.7)
<i>S. rodhaini</i>	Burundi	2002	11.51	28	11246613	11772768	10948674	8102216	404373 (5.0)

Table 1. Global isolates of *S. mansoni* sequenced and number of variants.

Number of reads mapped to the *S. mansoni* reference genome v5.2, median coverage attained and SNPs called by mpileup (MP), UnifiedGenotyper (UG), HaplotypeCaller (HC) and the consensus result (Overlap). The numbers of heterozygous SNPs from the consensus results are shown. The year column indicates the date of first passage.

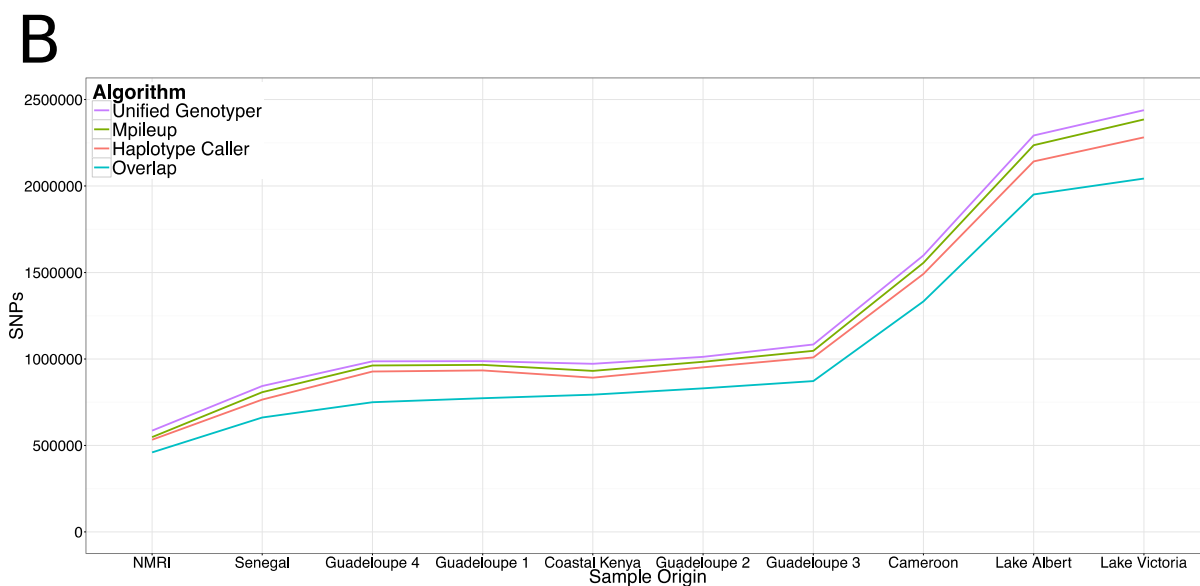
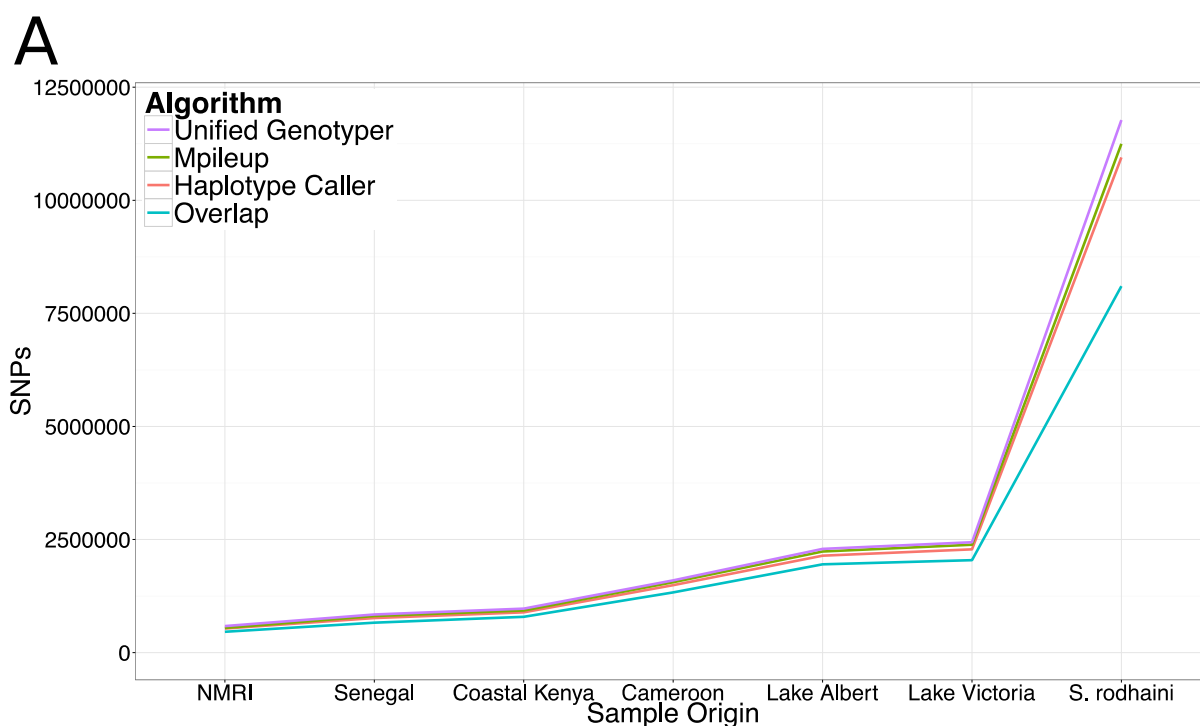


Figure 1. Number of SNPs when samples called against *S. mansoni* reference V5.2 by calling algorithm.

Consensus result (overlap) obtained by taking only those SNP sites called by all 3 callers A) shows results for all samples (*S. mansoni* & *S. rodhaini*) B) shows results only for *S. mansoni* samples

Population structure of S. mansoni

The dataset of filtered SNPs were used to understand the broad population genetic structure of *S. mansoni* and their relationship to *S. rodhaini*. A maximum likelihood phylogeny (Figure 2) confirmed that *S. rodhaini* was highly divergent from the ten *S. mansoni* samples. Rooting the phylogeny with the *S. rodhaini* sample as the out-group showed that the deepest split within *S. mansoni* separated Uganda (Lake Victoria and Lake Albert, bootstrap support (BS)=100) from the other populations, followed by Cameroon (BS=100). The rest of the sequences fell within a broader clade that included the Senegal, Coastal Kenya and all of the New World samples, suggesting a more recent divergence.

The phylogenetic pattern is supported by the principal component analysis (PCA) of SNPs (Figure 3). Within the *S. mansoni* sequences, Uganda and, to a lesser extent, Cameroon samples are separated from the others by the first principal component (PC), which accounts for 36.0% of the variance. Isolates within the more recently formed clade are separated along the second principal component axis (13.7% of variance), which distinguishes the two African samples in this group (Senegal and Coastal Kenya) from the Guadeloupe and NMRI (Puerto Rico-derived) samples. As expected, incorporating *S. rodhaini* in a PCA separates this sample from all of the *S. mansoni* isolates by the first PC, which accounts for > 70% of the variance and collapses the structure in the Senegal/Coastal Kenya/Guadeloupe group as a second PC only separates the Uganda and Cameroon *S. mansoni* from all other samples (Figure 3).

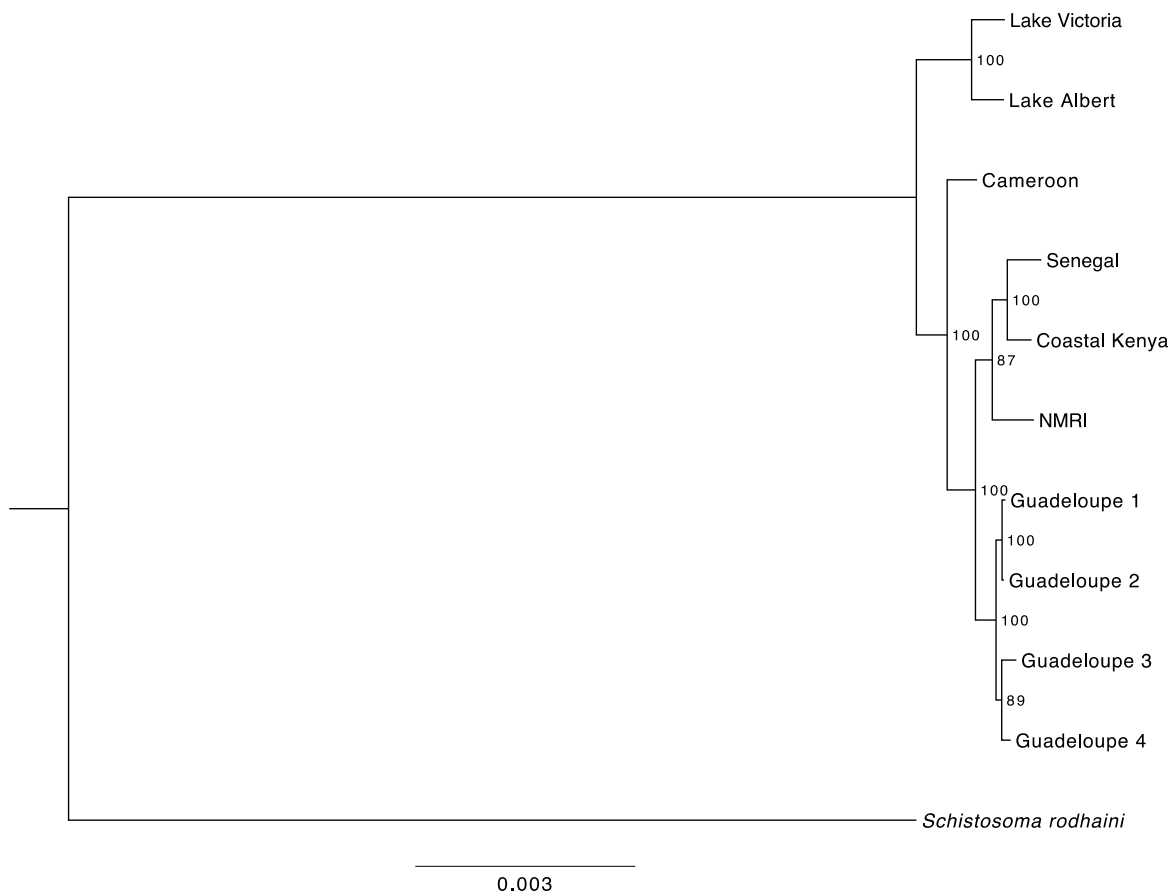


Figure 2. Maximum likelihood phylogeny of *Schistosoma* isolates based on whole-genome SNP data.

Values on nodes are the number of bootstrap replicates supporting the split induced by that node. *Schistosoma rodhaini* has been set as the out-group.

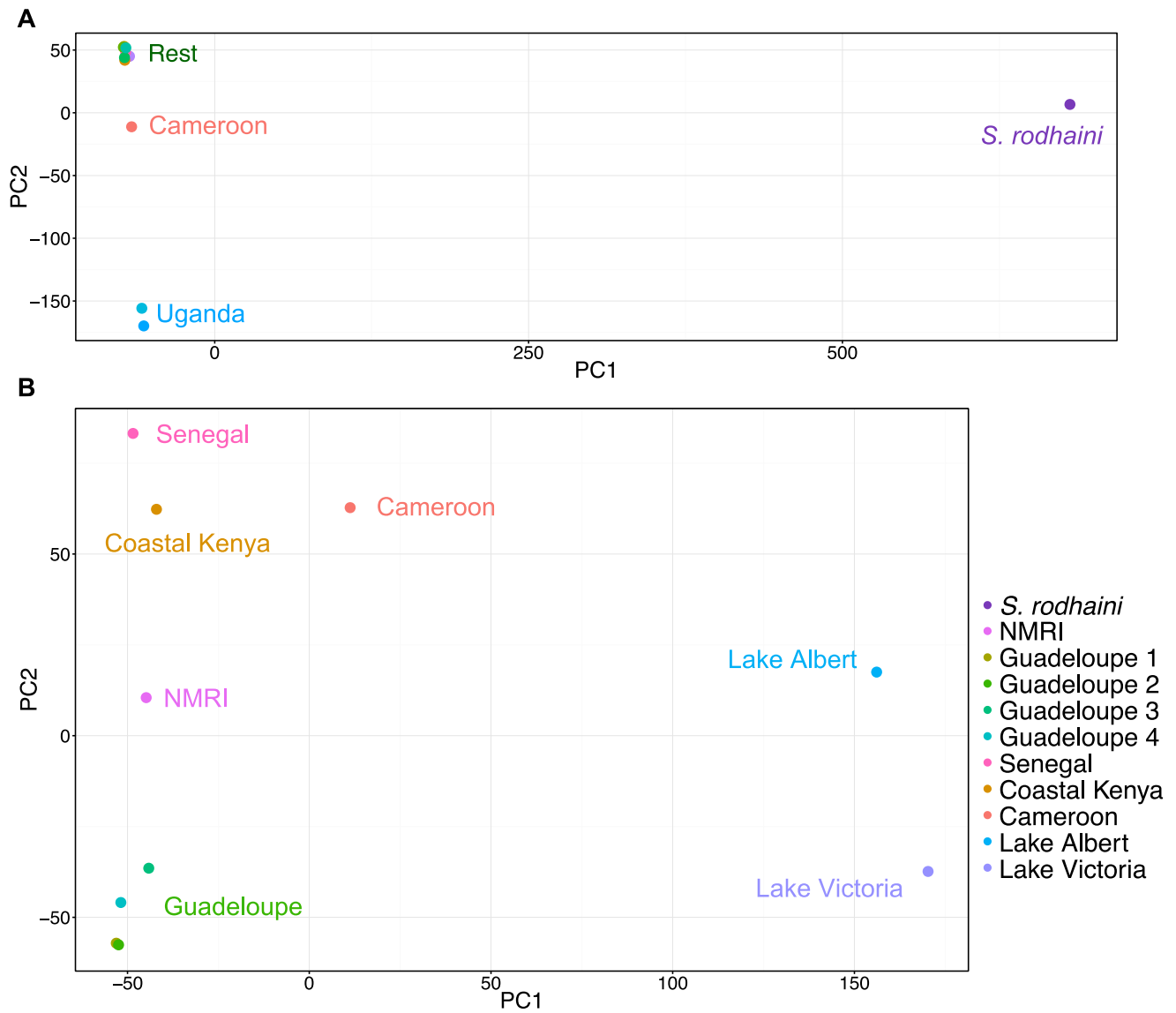


Figure 3. Principal component analyses of global isolate SNP data.

Panel A includes both *S. mansoni* and *S. rodhaini* samples (variance represented by PC1: 72.5 %, PC2 9.9%) and panel B includes *S. mansoni* samples only, showing more details of the relatedness of isolates within this species (variance represented by PC1: 36.0%, PC2: 13.7%)

Non-synonymous and synonymous changes

Tests for selection have identified interesting genes likely to be involved in host interaction in other parasites such as erythrocyte binding proteins in *P. falciparum* and *P. vivax* [146, 171], therefore this whole-genome dataset provides a unique opportunity to identify genes that are potentially host interacting in *S. mansoni*. I examined positive and negative selection between African *S. mansoni* and *S. rodhaini* by comparing d_N/d_S using a branch model and balancing selection within *S. mansoni* using the Hudson-Kreitman-Aguade and McDonald-Kreitman tests (see below). The ratio of d_N/d_S values across the filtered set of 9514 coding sequences was calculated with codeml using a branch model (see Methods), where the branch groups were set as African *S. mansoni* and *S. rodhaini* using the phylogenetic relationship in Figure 2.

The branch d_N/d_S values were centered around 1 (interquartile range: 0.44-1.1), examining the tails of the distribution I found 767 coding sequences showing evidence of positive selection ($d_N/d_S > 2$) and 2167 genes under strong purifying selection ($d_N/d_S < 0.1$), the distribution of these values is shown in Figure 4. Detailed functional annotation is lacking for most genes in *S. mansoni*, nonetheless I explored whether it was possible to detect any functionally enriched groups. Statistical overrepresentation analysis of the genes under positive selection showed an enrichment of gene ontology (GO) terms relating to the extracellular matrix (ECM) structural protein (GO:0005201, fold enrichment > 5 , $p = 2.42 \times 10^{-3}$). These five genes (Smp_197370, Smp_159600, Smp_170340, Smp_196840, Smp_033250) each contain multiple pfam domains of collagen triple helix repeats. All five genes have corresponding RNA-seq evidence of a significant increase in expression during the 24-hour schistosomula stage of the life cycle in pairwise comparisons with both the 3-hour schistosomula and the adult worms [124], implicating these genes in the adaptation of the adult worm stage to the definitive human host. I formally tested whether d_N/d_S values differed from 1, where coding sequences are assumed to be under drift rather than positive or negative selection (see Methods). Across all coding sequences, 52 were found to be significant ($p < 0.05$) of which 13 were under positive selection and 39 under purifying selection. Again, poor functional annotation limits to extent to which I could interpret these results, as 46 out of 52 significant coding sequences were unclassified. Despite this, I find evidence of significant positive selection on a number of genes involved in host-parasite interactions such as venom allergen-like (VAL) 21 gene (Smp_159290) and cercarial elastase gene

(Smp_119130). Both of these genes are expressed in the cercarial (mammalian infective) stage of the lifecycle and are known to be host-interacting [172, 173].

The d_s values on the *S. rodhaini* branch ranges from 0 to 0.82 with a median value of zero and only two coding sequences with tree length for $d_s > 0.5$, suggesting that saturation of synonymous mutations has not affected the results. Furthermore, methods based on modeling the codon substitution process have been shown to be more robust to saturation than alternative methods[174, 175].

Intra-species selection: Hudson-Kreitman-Aguade and McDonald-Kreitman tests

The Hudson-Kreitman-Aguade (HKA) test measures the ratio of polymorphisms (π) to fixed differences (K) across genes, whereby a higher HKA ratio (HKAr) reflects an enrichment of intra-specific variation that may be indicative of balancing selection or weak purifying selection. The mean value of K is 17.12 (pairwise nucleotide differences per kb) and the mean HKAr across genes is 0.132. The distribution of HKAr values is positively skewed with the majority of genes showing low values indicative of higher levels of fixed differences to polymorphisms (upper quartile = 0.15), as would be expected as the *S. rodhaini* outgroup is highly diverged. Therefore analysis of the few genes with high HKAr (> 0.25) may be indicative of balancing selection (see Figure 5 and below).

The McDonald-Kreitman skew (MKS) tests for an enrichment of non-synonymous mutations in either fixed difference indicative of directional (positive) selection ($MKS < -1$) or polymorphisms, indicative of balancing selection ($MKS > 1$) as measured by a 2x2 contingency table (see Methods). The genes were filtered to retain sequences where the absolute numbers of polymorphic or fixed differences were ≥ 5 and where the mean depth did not exceed 1.75 times the mean depth across all genes, so as to exclude possible duplicates ($n = 4174$). Under Fisher's exact test, 271 of the filtered genes had a p value < 0.05 . Of these sequences, 31 showed evidence of directional selection and 237 genes showed evidence of balancing selection. Gene ontology analysis among the 31 genes under directional selection showed that no terms were significantly over-represented among these genes when a Bonferroni correction was applied; similarly no pfam domains were significantly enriched among these genes. Of the 237 genes showing evidence of balancing selection, 3 GO terms were significantly over-represented; cell junction protein (fold

enrichment > 5 , $p = 1.13 \times 10^{-2}$), cell adhesion molecule (fold enrichment = 4.51, $p = 1.59 \times 10^{-3}$) and G-protein modulator (fold enrichment = 3.69, $p = 1.24 \times 10^{-2}$). Pfam enrichment of the positive MKT genes revealed immunoglobulin I-set domains to be enriched (PF07679, OR = 3, $p = 0.02$) and four dynein protein domains (PF03028, OR= 9, $p = 0.009$; PF08393, OR= 8, $p = 0.01$; PF12777, OR = 9, $p = 0.009$; PF12780, OR = 9, $p = 0.009$). When the HKAr and MKS distributions are overlapped, four genes are shown to have extreme scores in both tests (Figure 5), included in these are heat shock 40 protein (*hsp40*), a homeobox protein, Elongator Complex Protein and an uncharacterised protein.

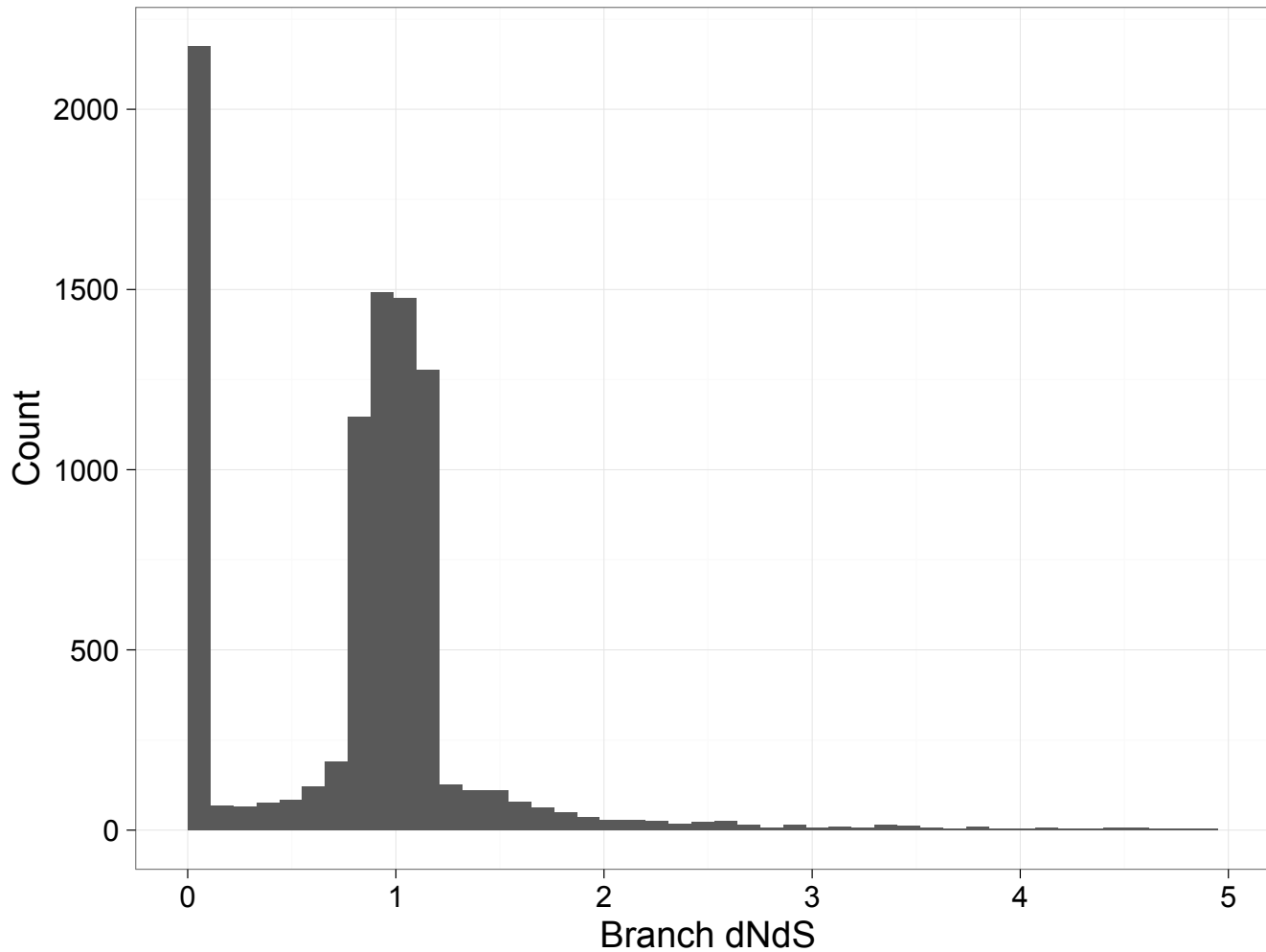


Figure 4. Histogram showing d_N/d_S values for 9514 coding regions in *S. mansoni* when compared against *S. rodhaini* using a branch model in codeml.

The majority of coding regions show $d_N/d_S \approx 1$, indicating neutral evolution. Values < 1 indicate coding regions under purifying selection and values > 1 indicate coding regions with variants under positive selection

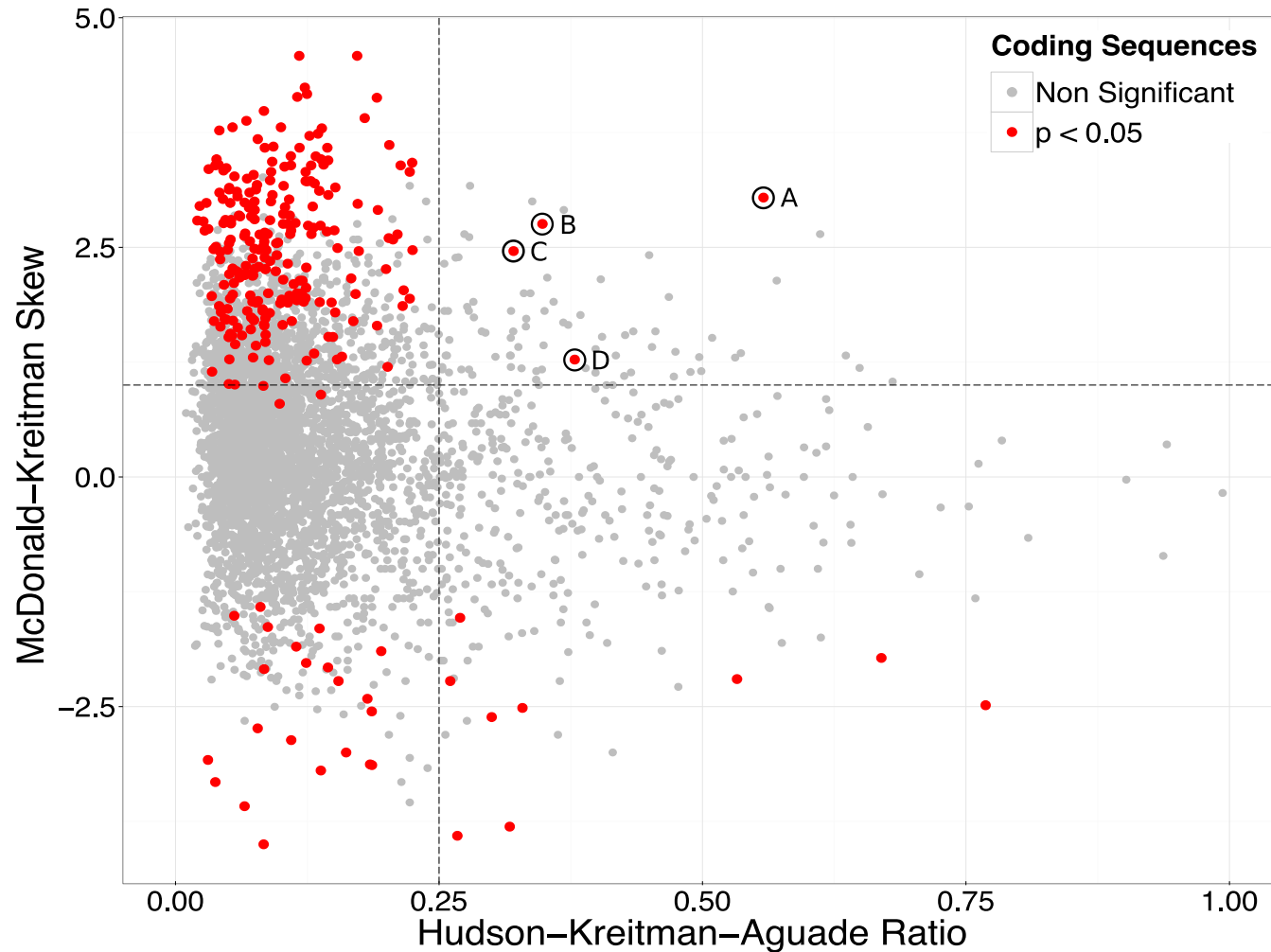


Figure 5. Population genetic evidence for balancing and directional selection on *S. mansoni* genes.

McDonald-Kreitman Skew (MKS) and Hudson-Kreitman-Aguade ratio (HKAr) performed on 4147 coding sequences. Genes with a significant MKS (Fisher's exact test, $p < 0.05$) are shown in red. Cutoffs of HKAr > 0.25 and MKS > 1 are used to identify genes under balancing selection (hashed lines). Significant genes with a high HKAr and positive MKS are A) Smp_035200- hsp40 protein B) Smp_157450- uncharacterised protein C) Smp_145220- Homeobox protein D) Smp_160080- Elongator Complex Protein 6

Historical demography of S. mansoni

Based on evidence of population structure from the phylogenetic analysis and PCA, along with an inference of the mutation rate, generation time and size of linkage blocks (Figures 2, 3, 6, 7 and Methods) and I constructed a putative population history for the samples using a number of coalescence models implemented in G-PhoCS (Figure 8, Methods). African *S. mansoni* and *S. rodhaini* were specified as two populations (Figure 8) to estimate the divergence date between species, and effective population size (N_e) before and after speciation. These posterior values are derived from coalescent models and scaled by two independently estimated parameters; mutation rate and generation time. The estimates of these parameters are discussed in the Methods and the Discussion.

Analysis of the results from this model shows that the split between African *S. mansoni* sequences and *S. rodhaini* occurred 126.5 thousand years ago (KYA) (95% CI 107.5-147.6KYA), see Figure 9 for the posterior distribution of the estimate of the time to MRCA when migration is permitted from *S. mansoni* into *S. rodhaini* populations. This introgression is unidirectional and appears to have taken place soon after the initial divergence between *S. mansoni* and *S. rodhaini*, though at a low frequency (proportion of *S. rodhaini* population arising via migration from *S. mansoni* per generation = 5.53×10^{-7}). Ancestral N_e before speciation ($N_e = 4.25 \times 10^5$) is shown to be more than four times greater than present N_e values for the extant African *S. mansoni* population as a whole ($N_e = 1.02 \times 10^5$) and more than 10 times greater than the N_e of extant *S. rodhaini* ($N_e = 1.50 \times 10^4$). No migration can be seen between *S. rodhaini* and any extant populations of *S. mansoni* in the models, which contrasts with recent evidence of hybridisation between natural populations of the two species in East Africa [108, 176], and suggests that these events have not occurred at a high enough frequency over the past 126.5KYA to be detected by the models.

The split between New World (Guadeloupe) and West African (Senegal and Cameroon) samples was estimated as 543 years ago. Given that the Guadeloupe samples were isolated from the field in 1983, this is dated to 1440 AD (95% CI 1118-1743 AD). The historical date of the arrival of slaves to Guadeloupe, 1669 AD [177], is therefore spanned by the 95% confidence interval. Different permutations of this model (for instance modeling the split between New World and only the Senegal strain and the inclusion of migration bands) produced near-identical results.

Additional models were used to calculate the divergence between East (Lake Victoria and Lake Albert) and West (Senegal and Cameroon) African samples into different populations. On the basis of the phylogenetic tree (Figure 2) and G-PhoCS evidence, I model the process with the Ugandan isolates as the basal clade and Cameroon and Senegal splitting off (Figure 8). The split between Uganda and Cameroon is dated to 5.92KYA (95% CI 4.71-7.16KYA). The N_e of the ancestral population shows no significant difference with Uganda ($N_e = 1.11 \times 10^5$ for both populations), whereas the Cameroon N_e showed a fourfold reduction since the split ($N_e = 2.44 \times 10^4$). Senegal split at a later date from the Ugandan isolates, 1.52KYA (95% CI 0.439-2.59KYA) and the effective population declines nearly tenfold ($N_e = 1.44 \times 10^3$). In both scenarios there is no evidence of migration in either direction following the population split. The N_e estimates of the G-PhoCS models are summarised in Table 2.

Effective population size estimates from this coalescent model assume a steady-state population existing along each branch of the population tree, and so cannot capture fine-grained demographic changes. Much finer temporal resolution of N_e is obtained by the PSMC model, in which each sample gives an independent estimate of how N_e has varied between discrete time periods. Results of this model largely corroborate those from G-PhoCS, with a large population size before 120KYA, when *S. mansoni* and *S. rodhaini* alleles were present in the same ancestral population (Figure 10; bootstrapping shown in Figure 11). The additional resolution of PSMC output reveals that a fall in N_e is then experienced by all sequences between 20-100KYA suggestive of a population bottleneck. A limitation of PSMC is that it estimates constant N_e within a time window, and so cannot differentiate between sudden and gradual reductions in population size and so the exact timing and rate of change in N_e is not possible to estimate exactly. In the last 7KYA N_e continues to decline in West African and New World populations, though N_e increases in the Lake Victoria and Lake Albert populations.

The PSMC results also corroborate some of the dates estimated by other methods. The demographic history of all isolates in the PSMC model follow the same trajectory until 7KYA ago (Figure 10), suggesting that *S. mansoni* existed as a single population until this point, approximately matching the estimate for the split between Uganda and Cameroon populations in G-PhoCS (95% CI 4.71-7.16KYA). The only deviation from this pattern is seen in the Senegal sample, which shows an increase in N_e 20-40KYA not present in the other isolates, though the

cause of this difference is not clear. The PSMC model cannot help understand contemporary changes in population size, while it has good resolution from 2-300KYA for the majority of sequences, more recently than 2KYA uncertainty increases as the number of coalescent events decreases, this is shown by the wide bootstrap values around the central tendency (Figure 11).

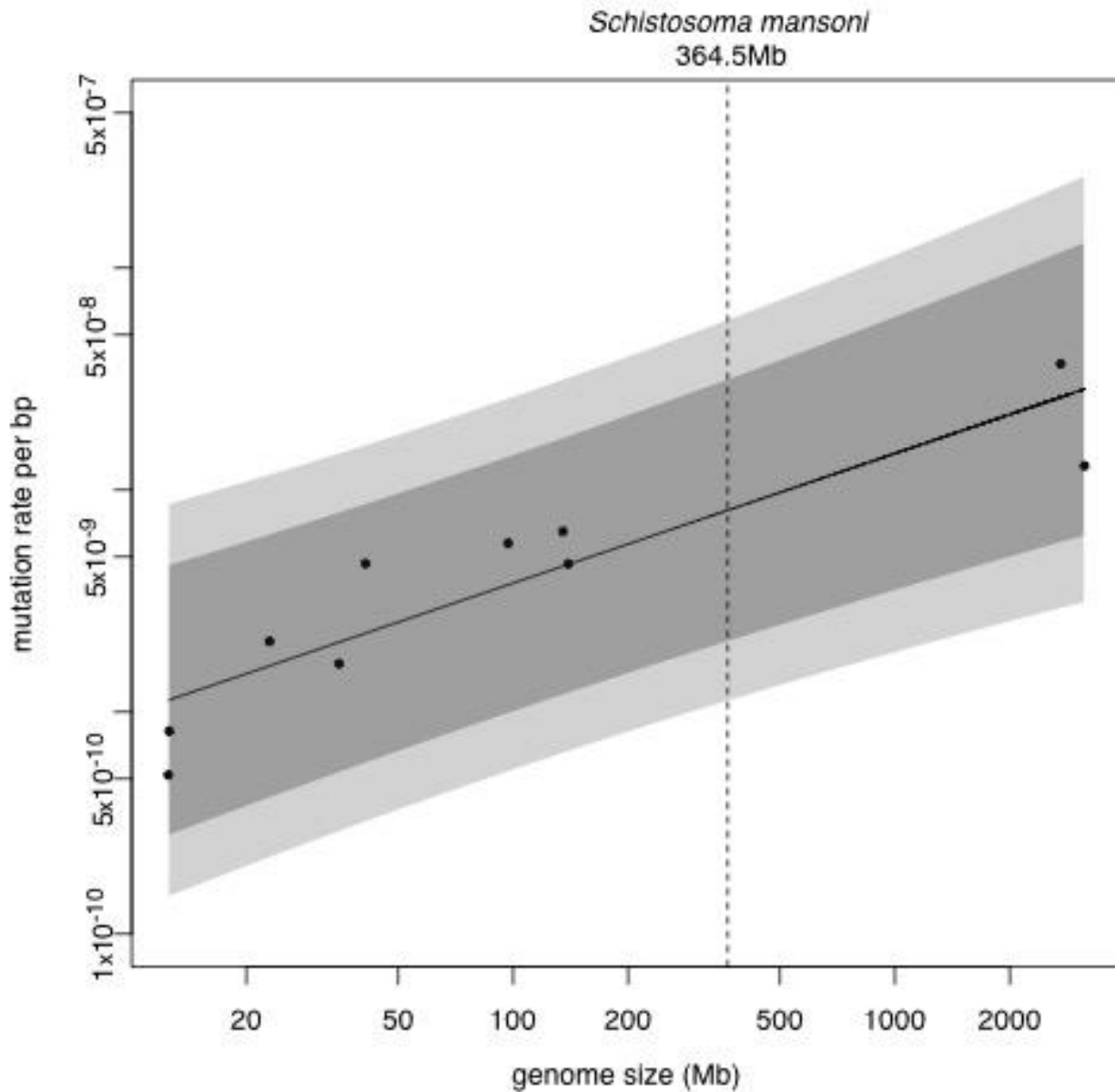


Figure 6. Estimate of the mutation rate (per basepair per generation) of *S. mansoni*.

Solid line indicates the slope of a linear regression between genome size in megabases (hashed line indicates the *S. mansoni* size) and mutation rate for known values for other organisms [169]; 95% and 99% confidence intervals are shown as dark and light grey bands respectively.

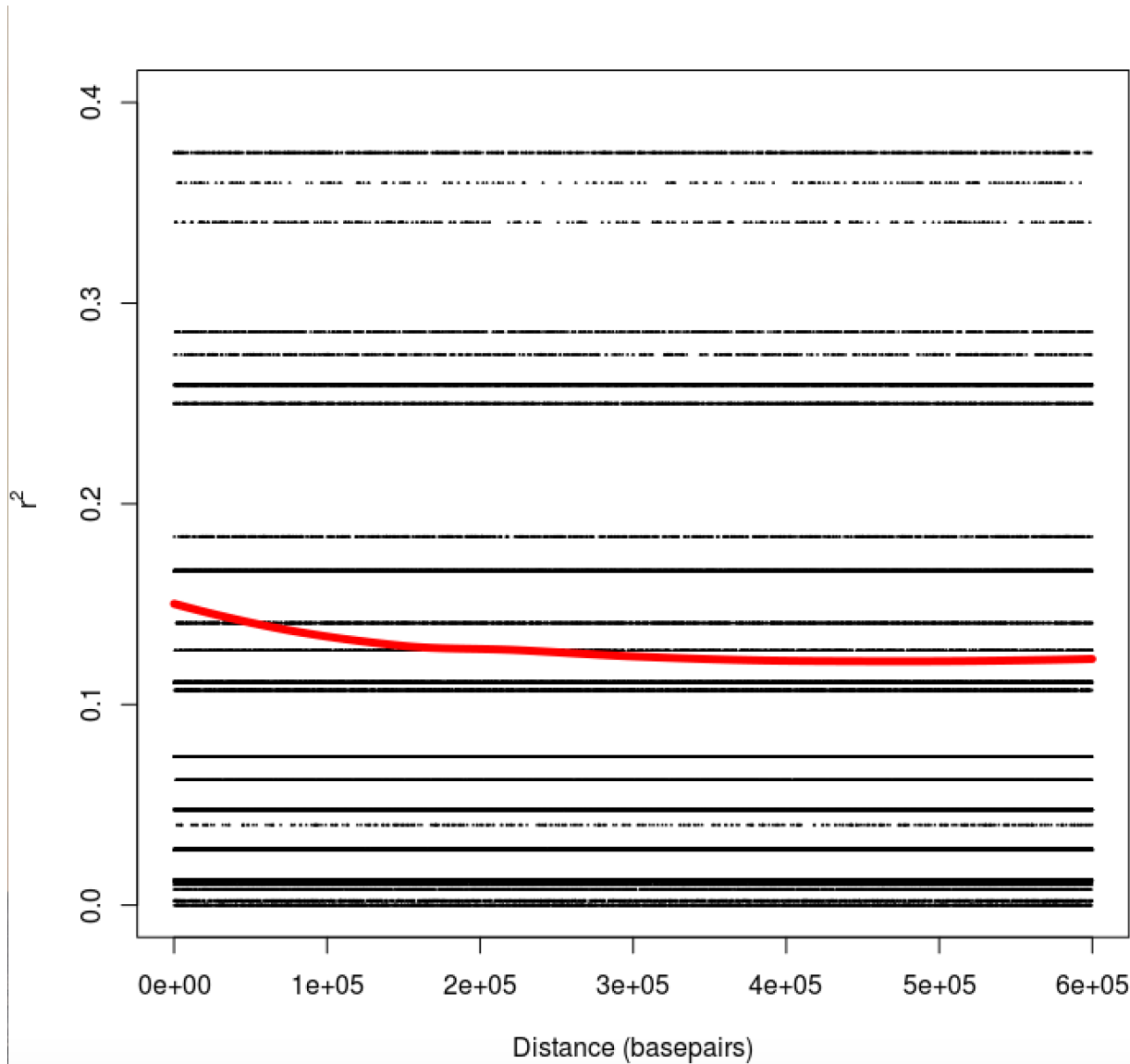
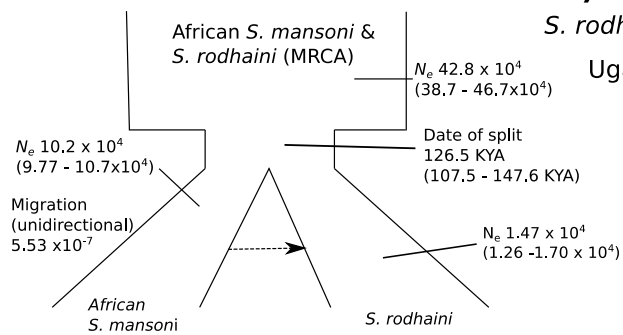


Figure 7. A measure of linkage disequilibrium (r^2) across the *S. mansoni* genome, calculated from phased SNP calls.

A loess line (red) is plotted to the data. The r^2 remains constant after a distance of approximately 400,000 base pairs, thus loci more than this distance apart are assumed to be in linkage equilibrium and evolve neutrally.

Model 1

S. mansoni - *S. rodhaini* split

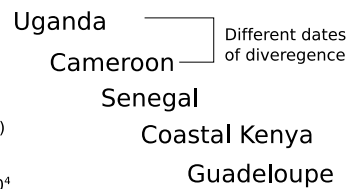


Model 3

Comparison of different branching models to validate phylogeny

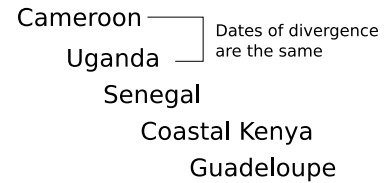
A)

S. rodhaini



B)

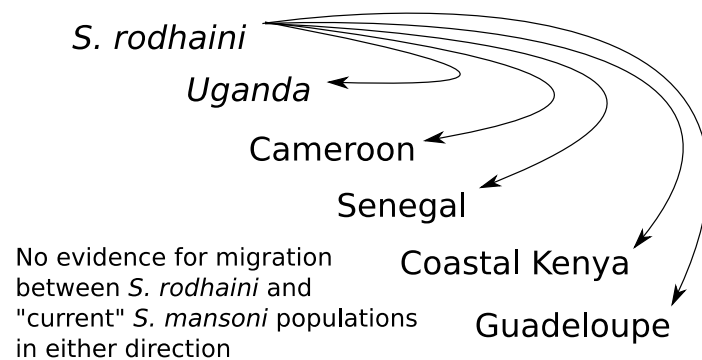
S. rodhaini



Model A is more robust as the dates of divergence are different for Uganda and Cameroon. This implies the branching order is correct, supporting the maximum likelihood phylogeny of Figure A

Model 2

Migration between *S. rodhaini* & *S. mansoni* populations



Model 4

West African - New World Split

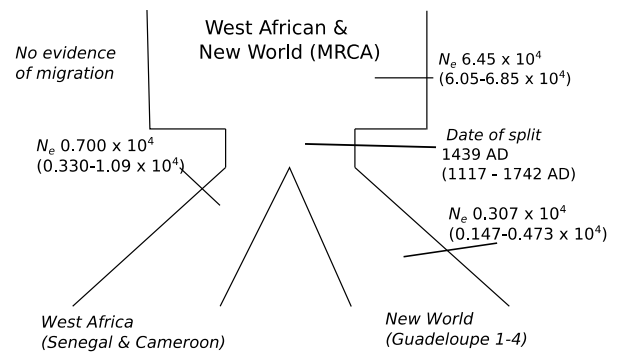


Figure 8. Models used for G-PhoCS coalescence simulations.

Parameters set as: Migration Rate Alpha = 0.02, Migration Rate Beta = 0.0001, Tau Initial = 0.005, Tau Beta = 20000

Population	<i>Mean N_e</i> (x 10⁴)	95% CI (x 10⁴)
LCA <i>S. mansoni</i> / <i>S. rodhaini</i>	42.8	38.7-46.7
Ancestral <i>S. mansoni</i>	10.2	9.77-10.7
Present <i>S. rodhaini</i>	1.47	1.26-170
Present East Africa (<i>S. mansoni</i>)	6.50	3.67-9.35
Present West Africa (<i>S. mansoni</i>)	0.700	0.330-1.09
Present Guadeloupe (<i>S. mansoni</i>)	0.307	0.147-0.473

Table 2. Effective population size estimates obtained from G-PhoCS.

Values shown are mean of the posterior probability distribution and upper and lower bounds of the 95% highest posterior density confidence intervals obtained by the model. Note that ‘present’ values are average of N_e along branches. All values are given to 3 significant figures.

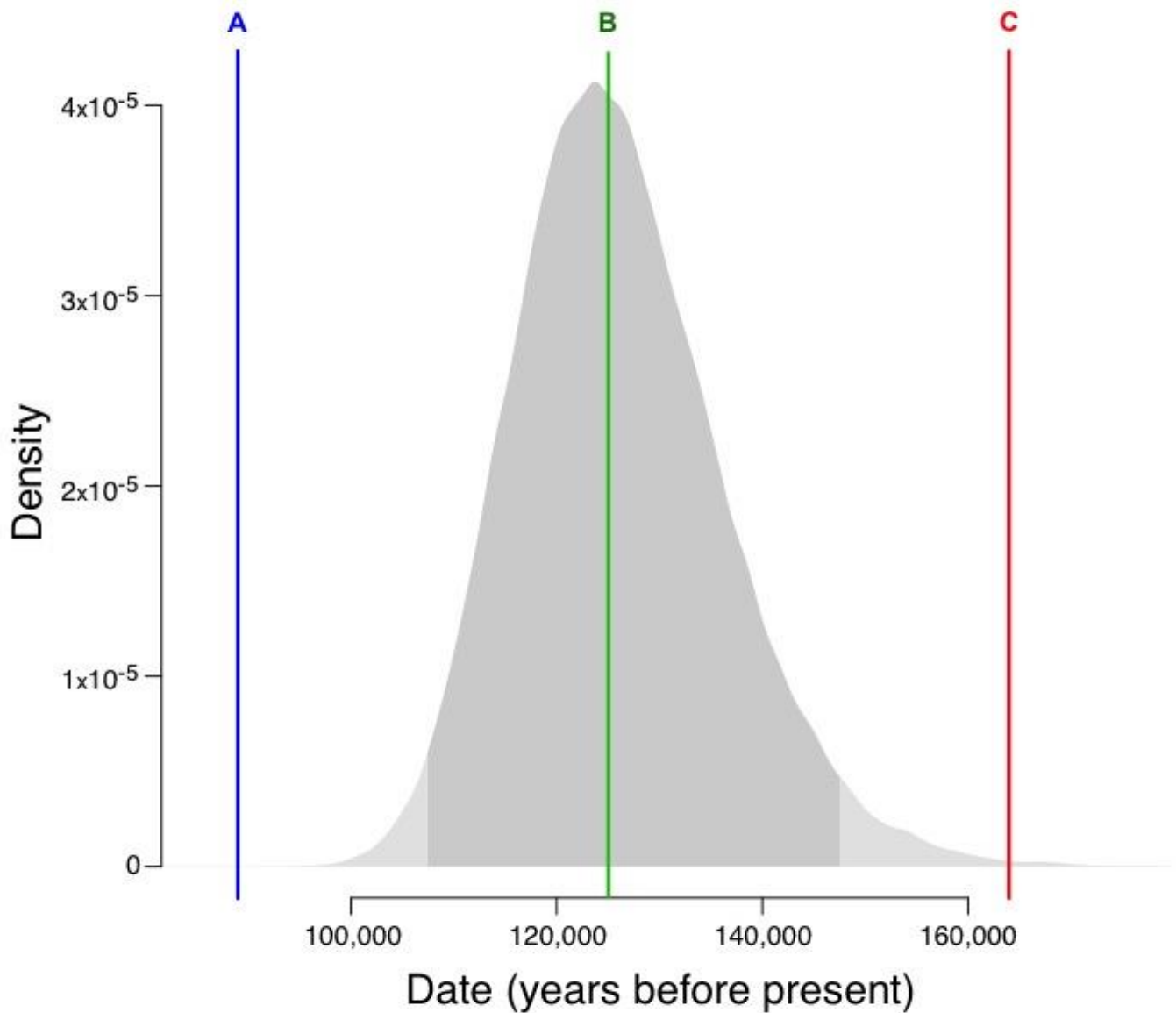


Figure 9. Date of *S. mansonii* – *S. rodhainii* divergence and archaeological evidence for fishing in Africa.

The posterior probability distribution for the date of the last common ancestor of *S. mansonii* and *S. rodhainii* from G-PHoCS (in grey), with the 95% highest posterior density confidence interval in darker shading. Coloured lines show central estimates for the dates of three early fishing sites in Africa, at (A) Upper Semliki Valley, Democratic Republic of Congo (dated to between 74 and 111KYA) [178], (B) Red Sea Basin, Eritrea (dated to between 118 and 132KYA) [179], (C) Mossel Bay, South Africa (dated to between 152-176KYA) [180]

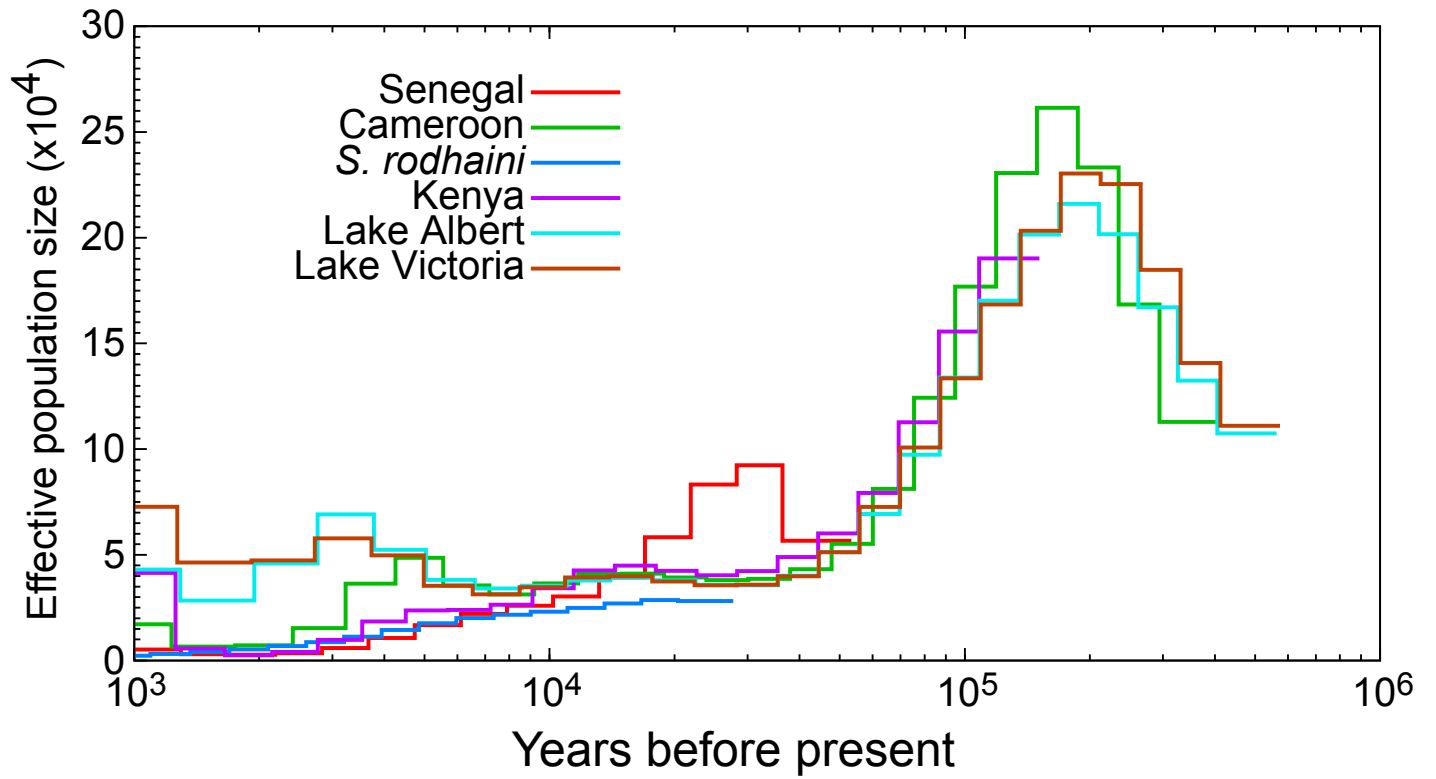


Figure 10. PSMC estimates of effective population size variation through time.

PSMC estimates of N_e are shown for all African *S. mansoni* isolates for dates between 1,000,000 – 1,000 years before present. Bootstrap confidence intervals for these lines are shown in Figure 11.

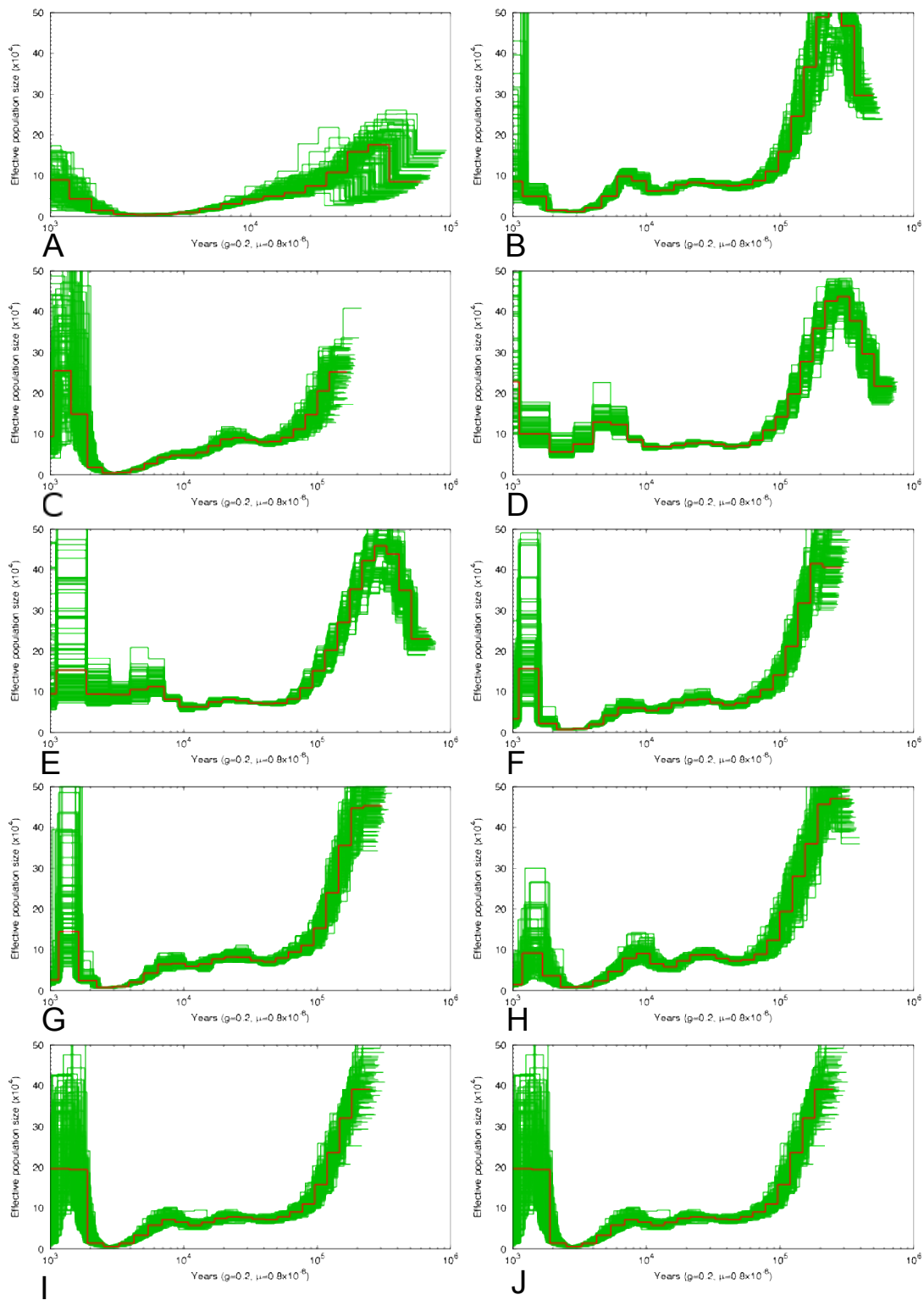


Figure 11. Bootstrap confidence intervals (100 replicates) around PSMC estimates.

Panels show isolates A) Senegal B) Cameroon C) Coastal Kenya D) Lake Albert E) Lake Victoria F) Guadeloupe 1 G) Guadeloupe 2 H) Guadeloupe 3 I) Guadeloupe 4 J) *Schistosoma rodhaini*

Discussion

The split between the New World (Guadeloupe) and West African populations (Senegal and Cameroon) is estimated to be between 1117-1742 AD. The two proposed possible timings of the Africa –New World migration are when the continents split 80-120 million years ago or during the Slave Trade of the 16-19th Centuries. The confidence interval calculated from the coalescence model spans the historical date of the earliest slaves to the New World (95% CI 1118-1743 AD). Slaves from French Colonies in West Africa were exported to islands in the French Caribbean, including Guadeloupe, from 1669 until 1864 and the total number of Africans taken from West Africa by French ships to Guadeloupe during this period numbered over 22,000 [177]. This result shows the data from whole-genome sequences to be consistent with both historical dating and the results found from previous studies on *S. mansoni* using smaller loci.

Genomic and archaeological evidence from other pathogens suggests that many human infectious diseases originated in the Old World and migrated to the New World [166], though the role of humans in transmitting diseases between continents is contested. For instance, *Mycobacterium tuberculosis*, which originated in Africa, was found in Peruvian remains dated to Pre-Columbian times. Based on genomic data, seals were implicated as the most likely reservoir that transmitted *M. tuberculosis* from Old to New World [181]. Diseases such as syphilis, caused by the spirochete *Treponema pallidum*, are assumed to have been transmitted to the New World with the Conquistadors of the 16th Century and this is supported by archaeological findings [182], though molecular data is lacking. The protozoan parasite *Leishmania chagasi*, which causes visceral leishmaniasis in the New World, is also thought to have been brought during the time of the Conquistadores though dogs imported into the Americas ~500 years ago, a migration event which caused a genetic bottleneck that has been observed in microsatellite loci [183]. The evidence presented here builds on theory that many human infectious diseases originated in the Old World and future population genomic studies of other pathogens have a valuable role to play in understanding the historical patterns of transmission that have shaped the current distributions of infectious diseases.

The evidence from coalescent models suggests that the emergence of *S. mansoni* as a distinct species occurred around 126.5KYA. This estimate is an order of magnitude later than previous estimates that placed the split at 2.8MYA and 1.2-10MYA respectively [20, 150]. I propose that this

most likely occurred in East Africa as the Ugandan isolates are the earliest diverging *S. mansoni* samples on the maximum likelihood tree. This supports the conclusions drawn from previous mtDNA barcoding studies [20]. *A priori* one might expect the origin of *S. mansoni* to be in East Africa as *S. rodhaini*, which shares a common ancestor with *S. mansoni*, has its present range in this part of the world. Furthermore, East Africa is where the earliest paleontological evidence for anatomically modern humans is found. A series of subsequent paleontological sites in this region indicate that early human populations inhabited this region throughout the late Pleistocene [184, 185].

I hypothesise that the adoption of fishing by human communities may have triggered the speciation of *S. mansoni* and *S. rodhaini* as this led to hunter-gatherer groups residing for longer periods by water bodies, leading to the establishment of *S. mansoni* transmission. Archaeological evidence from across Africa indicates a gradual accumulation in technology during the Middle Stone Age with the earliest evidence for exploiting aquatic resources from sites in South Africa, Eritrea and the Democratic Republic of Congo (DRC) [179, 180, 186, 187]. The earliest evidence is from Mossel Bay, South Africa where stone artifacts and substantial shellfish deposits are dated to 152-176KYA[180]. Similarly, at the Red Sea Basin in Eritrea deposits of stone tools and shells from marine organisms have been dated to 118-132KYA [179]. Early exploitation of freshwater resources is also seen in the Upper Semliki Valley, DRC where bone assemblages suggest that a well developed industry to process fish in Central Africa was established 74-111KYA [178, 187]. The posterior distribution for the speciation of the last common ancestor of *S. mansoni* and *S. rodhaini* overlaps with these archaeological dates (Figure 9), suggesting that changes to subsistence patterns in Africa around this time may have led to increased water contact and subsequently human infection with schistosomes.

The PSMC models show that the ancestors of all extant *S. mansoni* populations underwent a bottleneck between 20-100KYA. These changes may be driven by changes to human populations, the effect of climate on the freshwater habitats of the intermediate snail host, or a combination. Coalescence modeling of human populations in Africa support a decline in N_e from 50-100KYA, though this bottleneck is milder than in Non-Africans [153]. The East African climate was generally arid from 60KYA until the beginning of the Holocene [188] and fluctuations in temperature and precipitation throughout this period impacted on the size and depth of East

African Lakes, where the ancestral population of *S. mansoni* existed prior to 7KYA. Geological evidence suggests that Lake Tanganyika went through three phases of low lake levels at the end of the Pleistocene (35-40KYA, 23KYA and 18KYA) and Lake Victoria was almost completely dry between 12-17KYA. The lake levels all began to rise rapidly about 11KYA, reaching levels close to those observed in the present [189]. As PSMC is unable to specify exactly when in the 20-100KYA window the bottleneck in *S. mansoni* occurred, it is not possible to associate the fall in parasite N_e with any specific event. Therefore, climatic fluctuations, a fall in human N_e and changes to lake levels are all plausible explanations for the reduction in *S. mansoni* population size at the end of the Pleistocene.

By combining evidence from the phylogenetic tree with coalescence modeling I argue that *S. mansoni* originated in East Africa and that the ancestors of all extant populations remained there as a single population until 7KYA as PSMC output shows all populations follow the same demographic trajectory from 7-120KYA. More recently than this, the populations follow different paths as N_e rises in Lake Albert and Lake Victoria isolates and falls for New World and West African isolates. I hypothesise, therefore, this was around the time that populations of *S. mansoni* began to migrate across the continent and colonise areas of West Africa. This view is supported by evidence from G-PhoCS, which shows that Cameroon split away from the East African population 6KYA years ago, followed by the Senegal population 1.5KYA. Movement of human populations in Africa may explain the migration of parasites to new regions. The Bantu expansion was the most significant movement of peoples around Africa following the start of the Holocene followed by the Yoruba and Luhya populations expanding from 6KYA out of West Africa into central, southern and eastern regions [190]. Both the increase in human population size and the extensive movement of peoples across Africa following the milder conditions of the Holocene may have created conditions in which schistosome parasites could be spread through water sources as humans migrated across the continent. The population history of *S. mansoni* based on the models is summarised in Figure 12.

Both coalescence models presented here use two free parameters, generation time and mutation rate. I have made estimates of both these values as 0.2 years (see Methods) and 8.1×10^{-9} per basepair per generation (Figure 6) respectively [170, 191, 192]. Generation time, which includes the latency period in humans, the latency period in snails and the average time to transmit

between the two hosts, varies with water temperature of the snail intermediate host, the species of snail and the force of infection [23]. The definition of generation time is often controversial and is particularly complicated in the case of parasites with multiple hosts. Given the variability of such factors spatially and temporally, my estimate of the generation time is a simplification and may be subject to alteration on the basis of future experimental evidence [193]. Therefore, if revisions are made to either the mutation rate or generation time both dates of divergence and N_e estimates provided here will need revision. Despite these limitations, I find it encouraging that the estimate of the timing of migration to the New World overlaps with historical dating, suggesting that estimates of the free parameters are likely to be robust.

The Coastal Kenyan isolate shows a pattern in the phylogenetic tree and PCA significantly different to that expected given its distant geographical position from the Puerto Rican reference and proximity to Uganda. In all instances the Coastal Kenya isolate clusters with the New World and Senegal strains, as opposed to the other East African sequences. Previous phylogenetic analysis has shown that Coastal Kenya and Zambia form a distinct clade away from other East African parasites on the basis of a 451bp region of the *cox1* gene in the mitochondria [102]. This suggests that Coastal Kenya and Zambian *S. mansoni* have had a separate evolutionary history to parasites in the African great lakes, which may be due to separate histories of these regions. Throughout the 19th Century slaves were traded between Portuguese colonists along the Zambezi and Arab traders in the Horn of Africa predominantly via the port of Mombasa [194]. Thus a movement of people, and presumably parasites, occurred between these two regions, which may have contributed to their differentiation with other *S. mansoni* samples within East Africa. Denser sampling of the diversity of parasites in both East and West Africa will be needed to resolve the history of these parasites.

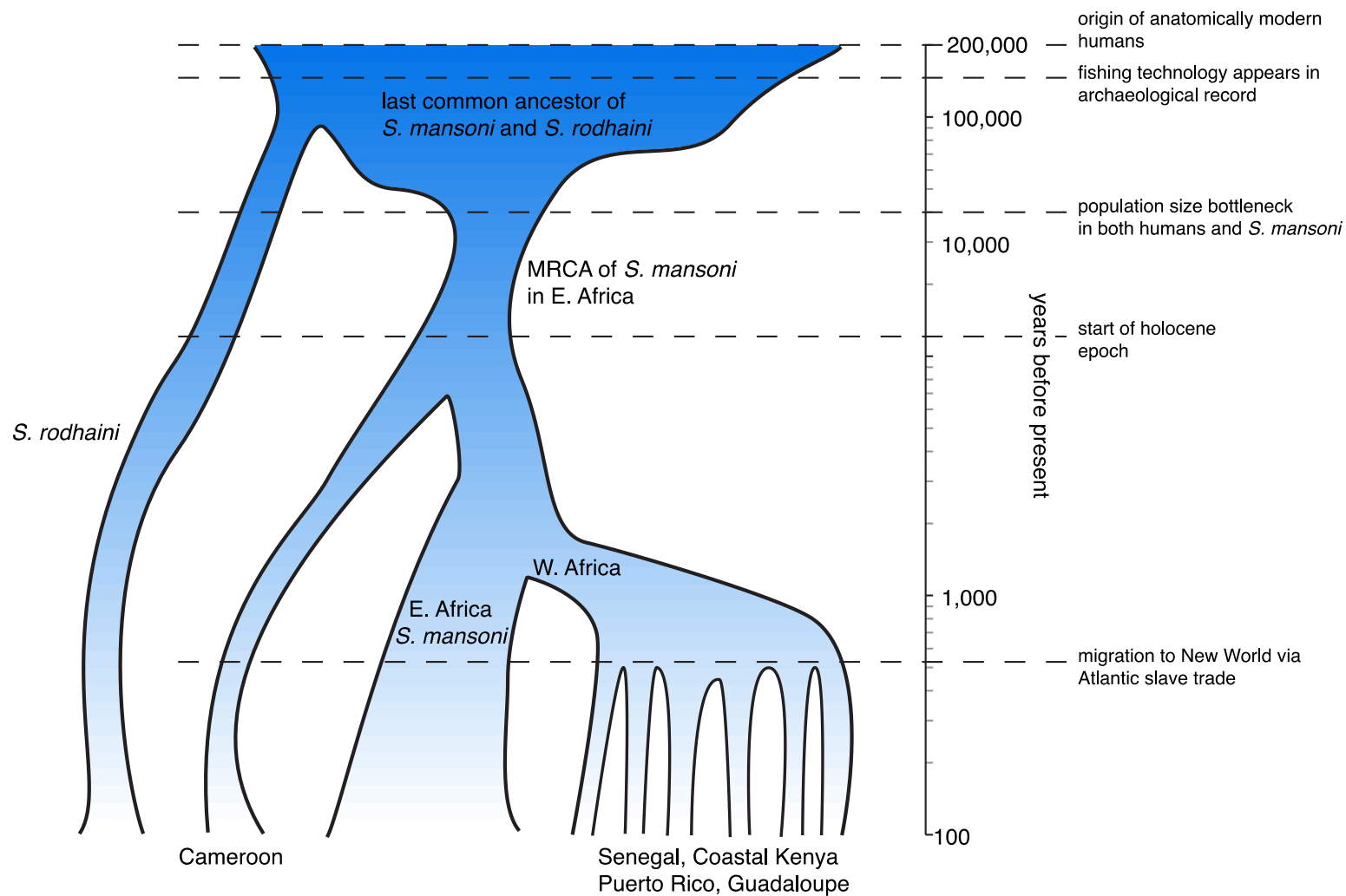


Figure 12. Summary of *Schistosoma mansoni* population history based on the evidence from this chapter.

Widths of branches represent approximate $\log N_e$ (effective population size. Note timescale is drawn on a log axis.

Through population genetic tests where *S. rodhaini* is specified as the out-group, genes have been highlighted that deviate from neutral expectations, either through an enrichment of non-synonymous substitutions or an over-representation of polymorphisms relative to fixed differences. Using a branch model of d_N/d_S , 767 genes were found showing evidence of positive selection ($d_N/d_S > 2$); of these a significantly enriched category of genes, ECM proteins, were found to have an increased level of expression during a period of host interaction in the schistosome lifecycle (24 hour schistosomula stage). This suggests that these genes may contribute to host invasion, as they have undergone rapid evolution as the parasite *S. mansoni* has become human-infective. When these genes were tested for significance under a likelihood ratio framework, VAL21 and a cercarial elastase gene were found to be under significant positive selection. The VAL genes excrete protein products that induce allergic responses. Transcriptomic evidence has shown that VAL21 is expressed primarily in the cercarial stage, suggesting it may play a role in invasion of the definitive mammalian host [172]. Cercarial elastase cleaves insoluble elastin, a major component of the dermis, and is the only protease required for mammalian host invasion [173]. While cercarial elastase genes are broadly conserved within the genus *Schistosoma*, the finding of a signal of positive selection in Smp_119130 may hold the key for human-specific invasion.

Genes with high values in the MKS, suggestive of balancing selection, showed enriched pfam domains in immunoglobulin I-set proteins, a domain linked to cell-cell interactions that could facilitate adaptation to the host. Four domains of dynein were also significantly enriched; this motor protein is involved in transport of organelles in cells, suggesting that the energetic requirement of *S. mansoni* cells is a target for selection [195]. As the isolates are not from first passage (see Methods), this represents a potential limitation on the tests for selection if laboratory passage has introduced novel selection pressures. This highlights the importance of using whole-genome sequences directly from clinical isolates in future studies.

Combining the MKS and HKA test revealed genes with extreme values, indicative of balancing selection. One of the main outliers is *hsp40*, part of a family of proteins known to regulate *hsp70s*, a gene family that encodes proteins that protect cells from stress and that have undergone an expansion in other platyhelminth species, such as tapeworms [196]. While there may be multiple explanations for the pattern of variation observed on any particular gene, the loci identified are subject to unusual evolutionary processes compared to the rest of the genome and so may provide

valuable targets for future functional experiments. As with most other non-model organisms, inferring the function of 'hits' from tests for selection can be difficult. Efforts are therefore underway to develop high-throughput functional genomic approaches to flatworm biology [197] and the genes that have been identified as outliers in the selection tests from this analysis will be fed into an ongoing RNAi screen to examine function at the Wellcome Trust Sanger Institute.

Chapter 3

Comparison of praziquantel efficacy against *Schistosoma mansoni* in Ugandan primary schools with different exposures to mass drug administration

Abstract

Mass drug administration (MDA) with praziquantel is the cornerstone of schistosomiasis control in sub-Saharan Africa. The effectiveness of this strategy is dependent on the continued high efficacy of praziquantel, however drug efficacy is rarely monitored using appropriate statistical approaches that can detect early signs of wane. I conducted a repeated cross-sectional study, examining children infected with *Schistosoma mansoni* from six schools in Uganda that had previously received between 1 and 9 rounds of MDA with praziquantel. I collected up to 12 *S. mansoni* egg counts from 414 children aged 6-12 before and 25-27 days after treatment with praziquantel. Individual patient egg reduction rates (ERRs) were estimated using a statistical model to explore the influence of covariates, including the number of prior MDA rounds. The average ERR among children within schools that had received 8 or 9 previous rounds of MDA (95% Bayesian credible interval (BCI) 88.23%, 93.64%) was statistically significantly lower than the average in schools that had received 5 (95% BCI 96.13%, 99.08%) or 1 (95% BCI 95.51%, 98.96%) previous round of MDA. I estimate that 5.11%, 4.55% and 16.42% of children from schools that had received 1, 5, and 8/9 rounds of MDA respectively had ERRs below the 90% threshold of optimal praziquantel efficacy set by the World Health Organization. The reduced efficacy of praziquantel in schools with a higher exposure to MDA may pose a threat to the effectiveness of schistosomiasis control programs and should be closely monitored.

Introduction

Interventions to control schistosomiasis have varied over time. Early national control programs, such as the Sudanese Blue Nile Health Project (started in 1979) and the Egyptian national program (started in 1988) used a combination of chemotherapy, snail control, provision of clean water and sanitation and infrastructural development [63, 64]. Such multi-approach control programs are now the exception rather than the rule [65]. Due to the low cost, safety and high efficacy of praziquantel, mass drug administration (MDA) has become the mainstay of national control programs.

Mass Drug Administration began in Uganda in 2003 and 400,000 treatments of praziquantel were distributed that year, with the support of the Schistosomiasis Control Initiative (SCI) [66, 67, 198]. This has since expanded to the treatment of approximately 1.5 million children annually [199], resulting in significant reductions in *S. mansoni* prevalence, intensity and associated morbidity [69, 71]. Similar successes have been reported in other countries implementing MDA, such as Burkina Faso [73]. Spurred by these results, the World Health Organization (WHO) revised their strategic plan from disease control to the interruption of transmission in certain African countries, including Uganda, by 2025 [70]. With the exception of certain islands implementing snail control [200], this ambitious plan remains solely reliant on MDA with praziquantel. Monitoring the continued efficacy of praziquantel is therefore paramount.

In the field, *S. mansoni* is commonly detected by the presence of eggs in stool, typically using the Kato-Katz thick smear method which is inexpensive and practical in low-resource settings [53]. As the test suffers from poor sensitivity, though close to 100% specificity [53], multiple readings improves the chances of detecting eggs in the stools of infected patients [55]. Egg counts measured before and after treatment are used to quantify drug efficacy as the egg reduction rate (ERR). The WHO has set a tentative ERR threshold of 90% as an indicator of optimal efficacy [82].

The sample ERR is typically calculated directly from egg count data by applying a simple formula (see Equation 2 in Methods). This is a population-level approach that is not readily compatible with individual patient data [62]. Such data are better analysed using mixed effects regression techniques that account for the correlation among parasite counts measured on the same individual or among counts measured from individuals within the same school. Failing to account

for the hierarchical (multi-level) structure of such data results in problems of pseudo-replication that can lead to erroneous estimates of parameter uncertainties. The mixed model framework includes both fixed and random effects that exert, among individuals sharing a set of covariates (e.g. individuals of the same age or sex), a constant or variable effect, respectively [19,20]. These methods are designed for longitudinally repeated measures and permit estimation of ERRs (and their associated uncertainties) for individual patients, which are useful for identifying poor responses and gradual shifts in anthelmintic efficacy [20].

Here I assess the effect of previous rounds of MDA on the efficacy of praziquantel against *S. mansoni* infections in primary schoolchildren in eastern Uganda. I explore whether the number of previous rounds of MDA is associated with efficacy of praziquantel, estimated as an ERR. Data are analysed from a repeated cross-sectional study of six schools that have received varying numbers of MDA rounds over a period of eleven years. A generalized linear mixed model (GLMM) implemented in a Bayesian framework is fitted to data on schistosome egg counts before and 25-27 days after treatment, estimating praziquantel efficacy at the levels of individual patients and schools.

Methods & Materials

Study Sites

The field team and myself (see Acknowledgments) conducted fieldwork in Mayuge and Tororo districts, Uganda from May – July 2014 (Figures 13 and 14). Six government-run primary schools were included into the study, five from Mayuge district and one from Tororo (Table 3). None of the schools had piped running water or a pump and all sanitation facilities were drop pits. Approximately 6 months had elapsed between the previous round of MDA and the start of the present study. The MDA in Tororo in 2013 was the first MDA in this district and achieved only 21% coverage. MDA in Mayuge has been ongoing since 2003, although MDA did not take place in any schools during 2008-9.

Figure 13. Maps of study sites in Uganda.

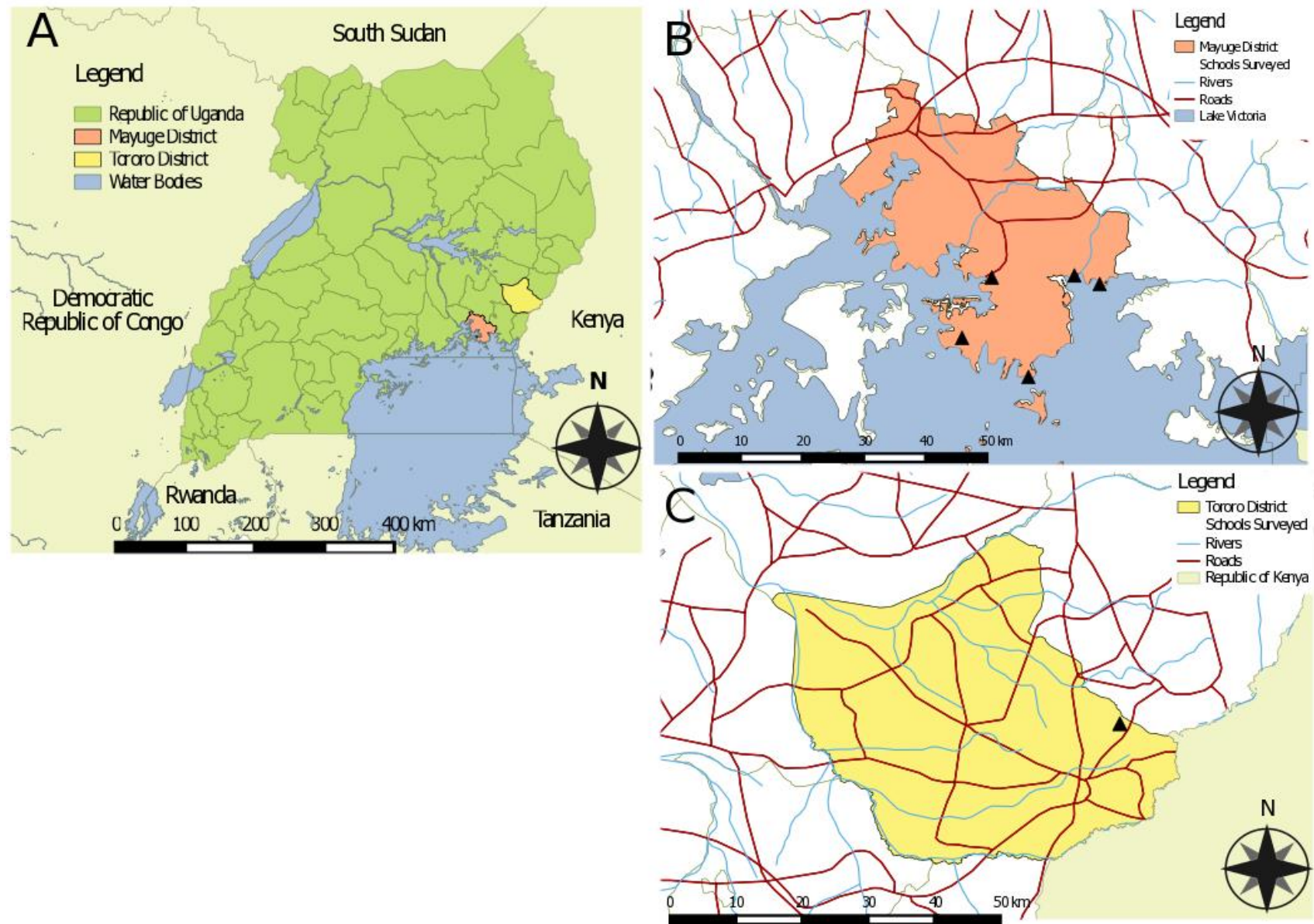


Figure 14. Photos taken during recruitment in Mayuge District, Uganda.

A



B



Figure 13. Maps of study sites in Uganda.

Panel A is a map of Uganda in an East African context with Mayuge and Tororo districts highlighted. Panel B shows the location of the schools surveyed in Mayuge district. Panel C shows the location of the school surveyed in Tororo district.

Figure 14. Photos taken during recruitment in Mayuge District, Uganda.

Panel A shows children lining up for recruitment, holding newspaper and toilet roll to provide a stool sample. Panel B shows Kato-Katz reading and preparation and children waiting to be treated in a classroom. Photos taken by Thomas Crellen.

Table 3. Sample estimates of the prevalence and intensity of *Schistosoma mansoni* infection in children from six primary schools in eastern Uganda before treatment with praziquantel.

The number of previous rounds of Mass Drug Administration (MDA) the school has been exposed and corresponding MDA Exposure Categories are shown. Infection intensity is quantified as the arithmetic mean number of egg per gram of faeces (EPG). Confidence intervals (CIs) are calculated using a block bootstrap approach to account for correlation among egg counts repeatedly measured from the same child.

Table 3. Sample estimates of the prevalence and intensity of *Schistosoma mansoni* infection in children from six primary schools in eastern Uganda before treatment with praziquantel.

School	District	MDA ^a Exposure Category	Recruited	Tested at Baseline	Positive for <i>S. mansoni</i> (%)	Mean EPG ^b (95 % CI ^c)
Bwondha	Mayuge	High (9 rounds)	98	96	87 (90.63)	742 (518-973)
Bugoto	Mayuge	High (9 rounds)	183	170	144 (84.71)	436 (339-534)
Musubi	Mayuge	High (8 rounds)	128	127	120 (94.49)	465 (362-600)
Bukoba	Mayuge	Medium (5 rounds) ^d	120	118	53 (44.92)	92 (67-128)
Bukagabo	Mayuge	Medium (5 rounds) ^d	120	118	57 (48.31)	306 (197-442)
Kocoge	Tororo	Low (1 round)	120	120	81 (67.50)	347 (245-456)
Total			769	749	542 (72.36)	382 (338-446)

^a Mass Drug Administration ^b Eggs Per Gram ^c Confidence Interval

^d Estimated from district-level MDA coverage from Ministry of Health

Enrolment

At baseline 769 children were recruited from six primary schools. Children were randomly sampled within the age group 6-12 years and by sex. Children enrolled in the study were examined from 1-3 days before and 1-3 days after treatment with duplicate Kato-Katz slides per stool per day (i.e. a maximum of 12 counts per child, up to 6 pre and 6 post treatment). Treatment with praziquantel and albendazole, for soil-transmitted helminths (STH), was overseen by a nurse and children were observed for one hour following treatment to check for side effects and to exclude from the efficacy analysis any children who vomited within an hour. Children were followed up 25-27 days after treatment.

Inclusion criteria for the baseline analysis was that the children were present for at least one day before treatment and so provided a minimum of 2 Kato-Katz thick smears for diagnosis. This gave a sample size of $n = 749$ patients. Inclusion criteria for the efficacy analysis were that the child was i) present for at least 1 day before and 1 day after treatment; ii) positive for *S. mansoni* at baseline and, iii) successfully treated with praziquantel and albendazole without vomiting. This gave a sample size of $n = 414$ patients (Figure 15).

Sample Egg Reduction Rate (ERR)

The sample ERR is calculated using

$$\text{Sample ERR} = 1 - \frac{\text{mean egg count after treatment}}{\text{mean egg count before treatment}} \quad (2)$$

where the mean is calculated at the level of the school using all egg counts, rather than individual level averages. A non-parametric percentile block bootstrap method was used to calculate confidence intervals (CIs) associated with the sample ERR. Block bootstrap methods [201] account for correlation among observations (egg counts) from the same individual by randomly sampling (with replacement) blocks of data; here, all of an individual's egg counts before and after treatment. This method was also applied to calculate CIs associated with mean egg counts before and after treatment.

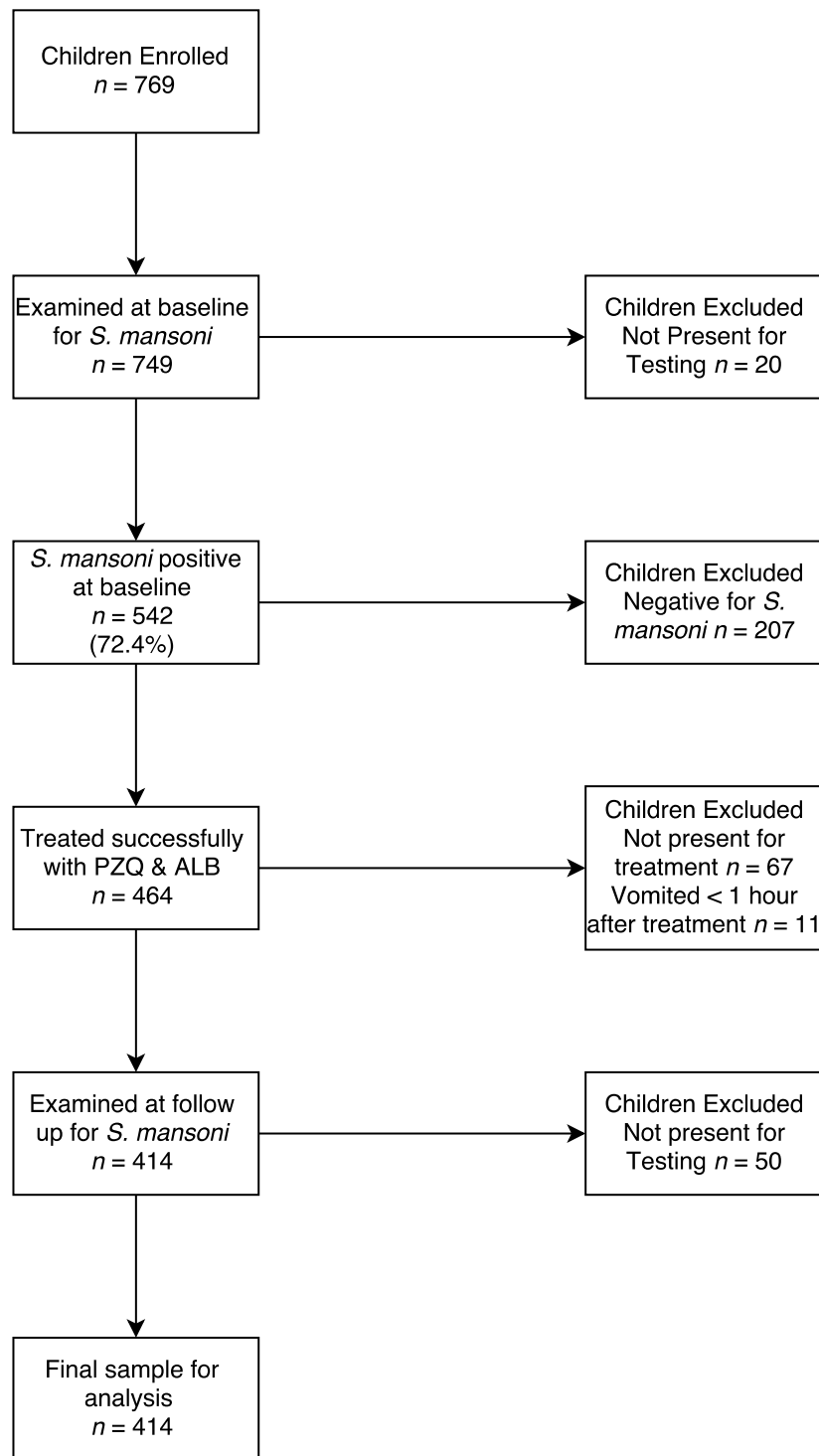


Figure 15. Flow diagram of patient recruitment and exclusion criteria for egg reduction rate analysis after treatment with praziquantel (PZQ) and albendazole (ALB).

Covariates

The schools were selected to give a variety of different MDA histories: one round for Kocoge; an estimated five rounds for Bukoba and Bukogabo; eight rounds for Musubi and nine rounds for Bugoto and Bwondha, which I classify as low, medium and high previous exposure to MDA respectively (Table 3). While school level MDA records were not available for Bukoba and Bukagabo, the number of MDA rounds were estimated based on information from Mayuge district officials and district-level MDA coverage from the Ugandan Ministry of Health. The sex and weight of each child were recorded at enrollment. Egg counts for STHs (*Ascaris lumbricoides*, *Trichuris trichiura* and hookworm species) were also collected. Weight was categorized as: low, <22kg; medium, 22-26kg and high, >26kg. Age and height were recorded, but dropped as covariates due to their correlation with weight. Co-infection with STH was defined as a binary variable.

Statistical Analysis

A statistical model fitted to the *S. mansoni* egg count data in a Bayesian framework using a recently method designed to accommodate egg count data before and after treatment [62, 84]. The distributional assumption of the model is that egg counts measured from a patient at a specific time (before or after treatment) are Poisson distributed, with an observation level random effect permitting extra-Poisson variation or overdispersion [19, 22, 23]. The key systematic component of the model is a log-linear regression structure that describes the change in egg counts after treatment, $x = 1$, compared to before treatment, $x = 0$, in a multiplicative fashion such that the accompanying treatment effect coefficient β quantifies the relative change and the ERR is given by $1 - \exp(\beta x)$. Covariates enter the model as fixed effects interacting with x , allowing ERRs to vary among patients of different sex, weight category and MDA exposure. The fixed effects formula is:

$$\text{egg count} \sim \text{MDA treatment category} * \text{time} + \text{weight category} + \text{sex} + \text{STH co-infection} \quad (3)$$

Random effects coefficients, specific to each patient, school and also interacting with x , permit ERRs to vary among patients and among schools sharing the same (fixed effect) covariate values. The correlation (ρ) between random effects (at individual patient or school levels) quantifies the association between egg counts before treatment and the treatment effect. A *positive* correlation at the individual level corresponds to individuals with higher than average egg counts at baseline tending to have *lower* ERRs and *vice versa* for a negative correlation.

We fitted the model to the data on egg counts in a Bayesian framework using Markov chain Monte Carlo techniques [62, 84, 202, 203]. We defined uninformative / vague parameter prior distributions and ran the Markov chains for 100,000 iterations with a 2,500 iteration burn-in, yielding an effective sample size of >1000 for all parameters and hyperparameters. We visually checked that the chains had converged to a stable posterior distribution. We constructed posterior distributions for ERRs associated with schools and MDA exposure categories by marginalizing over the appropriate subset of the coefficients. The analysis was performed using R version 3.0 [204] and the package MCMCglmm v2.2 [203].

Results

Summary epidemiological statistics

Baseline epidemiological data for each of the six schools are shown in Table 3. The analysis was performed on 749 children who were present for at least one day of sampling before treatment. The prevalence of *S. mansoni* at baseline in the schools surveyed ranged from 44.92% in Bukoba to 94.49% in Musubi. The arithmetic mean eggs per gram (EPG) was calculated by averaging over egg counts from schools and multiplying by a factor of 24, rather than from the fitted model, to be comparable with previous studies in Uganda [77]. The intensity of infection varied seven-fold among schools, from 92 EPG (95% confidence interval, CI: 67-128 EPG) in Bukoba to 742 EPG (95% CI: 518-973 EPG) in Bwondha.

Of the 414 children included in the praziquantel efficacy analysis (Figure 15), 209 (50.48%) were female and the age range was 6-12 years with a median of 8.5 years. A total of 4102 egg counts (approximately 10 per child) were collected from Kato-Katz smears: 2090 before treatment and 2012 after treatment. Co-infections of *S. mansoni* and STH were present in 141 patients (34.06%) of which the majority were hookworm (139, 33.57%). Comparing the 414 children included in the efficacy analysis with the 128 excluded children, there are no significant differences in the fixed effect characteristics (sex, age, weight, MDA exposure) between the two groups as measured by a chi-square test of association.

Variation in praziquantel efficacy among children

The estimated posterior distributions of the model coefficients associated with praziquantel efficacy are summarized in Table 4 as means and 95% Bayesian credible intervals (BCIs). The (Bayesian) p values [205] associated with each posterior distribution are also presented. Negative coefficients (or their exponent; < 1) associated with covariate-treatment interactions indicate a lower egg count after treatment and a higher ERR and *vice versa* for positive coefficients (or their exponent; > 1). Because the model is specified at the level of the individual patient, coefficients are interpreted as reflecting the fixed covariate effect on a *typical* child (i.e. when the random effects adjustments are equal to 0) [84].

The negative coefficients associated with egg counts measured after treatment from different low ($\beta = -1.32$, $p = 0.10$) and medium ($\beta = -1.42$, $p = 0.060$) MDA exposure categories indicates that a typical child within a school that has received fewer previous MDA rounds has a *higher* ERR than a typical child in a school that has received more previous rounds of MDA, though these effects are not statistically significant. The positive coefficients associated with medium ($\beta = 0.83$, $p = 0.0024$) and low ($\beta = 1.21$, $p = 0.048$) weight categories indicate that a typical child in these weight categories has a *lower* ERR than a typical child in the high weight category. The coefficients of other covariates fitted as interaction terms in the model (sex and co-infection with STH) are not statistically significantly different from zero. The estimated negative correlation between children's egg counts before treatment and the treatment effect ($\rho = -0.17$) indicates that children with higher egg counts tend to have *higher* ERRs, albeit the effect is of only marginal statistical significance (95% BCI: -0.32, -0.019).

Individual children's ERRs are variable within each school and MDA exposure category, however in the high MDA exposure category 16.42% (95% BCI: 13.42%, 19.78%) of the posterior mean individual ERRs are below the 90% reference efficacy [82] compared with 4.55% (95% BCI: 1.22%, 9.76%) and 5.11% (95% BCI: 0.00%, 10.94%) in the medium and low categories respectively. The BCIs quantify the considerable uncertainty in the individual mean ERR estimates and reflects both the variation among egg counts measured from the same child and the number of counts measured per child, which ranged from 4 to 12. The estimated ERRs and associated 95% BCIs for each child, stratified by school and MDA exposure category are shown in Figure 16, the change in estimated egg counts before and after treatment are shown for each child in Figure 17.

Table 4. Estimated coefficient posterior distributions of the generalized linear mixed model fitted to *Schistosoma mansoni* egg counts collected from 414 children from six primary schools in eastern Uganda before and after treatment with praziquantel.

The posteriors are summarised by their mean and 95% Bayesian credible interval (BCI). The coefficients estimates all interact with treatment and thus correspond to the effect of the covariate on the egg reduction rate (ERR). Negative coefficients (or their exponent; < 1) associated with covariate-treatment interactions indicate a lower egg count after treatment and a higher ERR and vice versa for positive coefficients (or their exponent; >1) (see Methods)

^a Bayesian Credible Interval ^b not applicable; ^c mass drug administration; ^d soil-transmitted helminth

Covariate	Category	Coefficient posterior mean (95% BCI ^a)	Exponent coefficient posterior mean (95% BCI)	Bayesian <i>p</i> value
Intercept	NA ^b	2.32 (1.36, 3.25)	10.79 (2.17, 21.19)	0.01
Treatment	NA	-5.65 (-6.81, -4.55)	3.91x10 ⁻³ (5.82x10 ⁻⁴ , 8.86x10 ⁻³)	<0.01
MDA ^c exposure	Low (1 round)	-1.33 (-3.20, 0.48)	0.328 (3.36x10 ⁻⁹ , 1.03)	0.10
	Medium (5 rounds)	-1.42 (-2.94, 0.25)	0.293 (9.34x10 ⁻¹³ , 0.793)	0.06
Weight	Low (<22kg)	1.21 (0.40, 2.03)	3.57 (1.12, 6.72)	<0.01
	Medium (22-26kg)	0.83 (0.01, 1.65)	2.45 (0.845, 4.74)	0.05
Sex	Male	0.13 (-0.48, 0.78)	1.19 (5.32, 1.98)	0.67
Co-infection with STH ^d	Positive	0.27 (-0.40, 0.99)	1.37 (0.569, 2.43)	0.44

Figure 16. Estimated egg reduction rates (ERRs) of *Schistosoma mansoni* from 414 children in eastern Uganda treated with praziquantel.

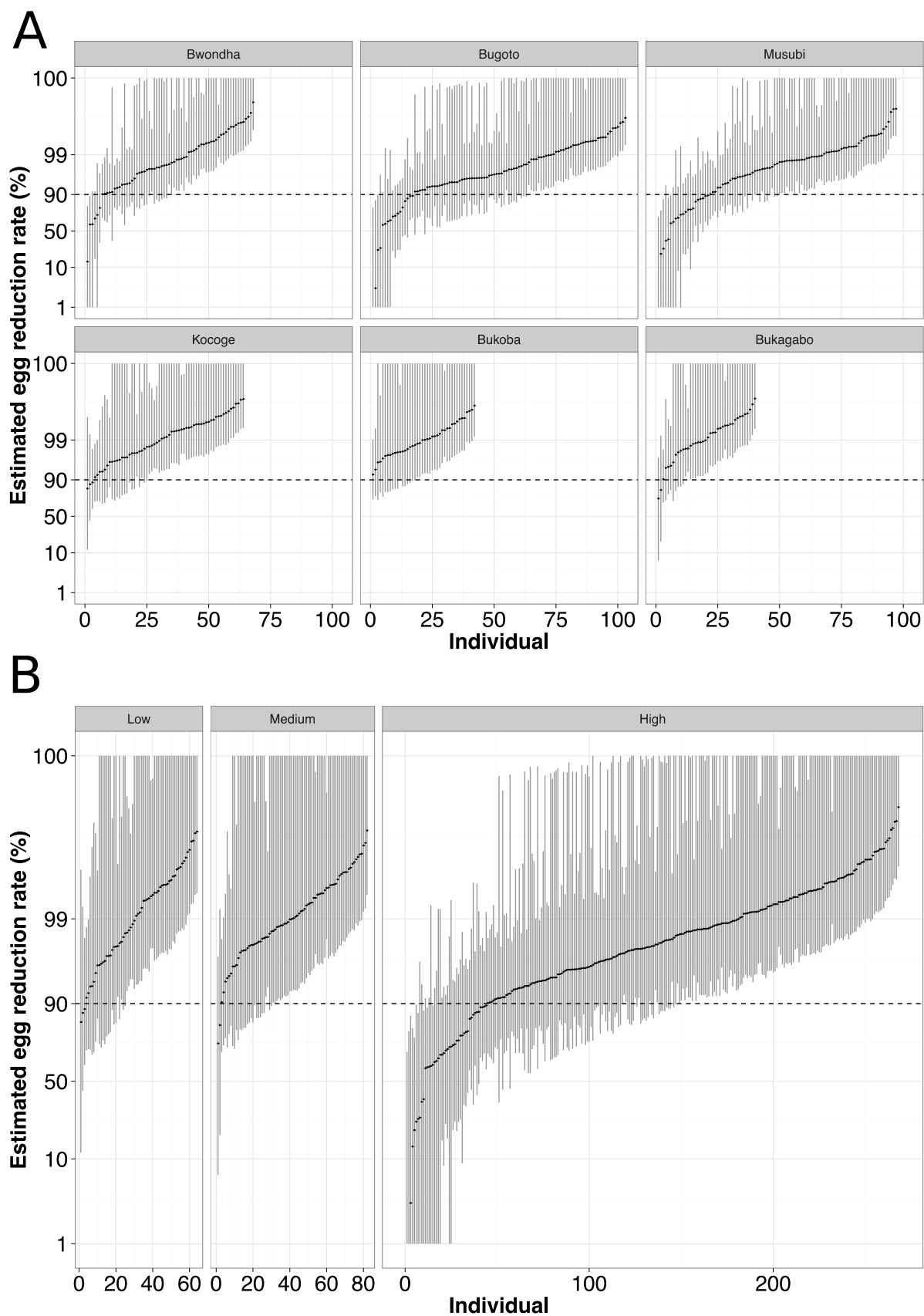


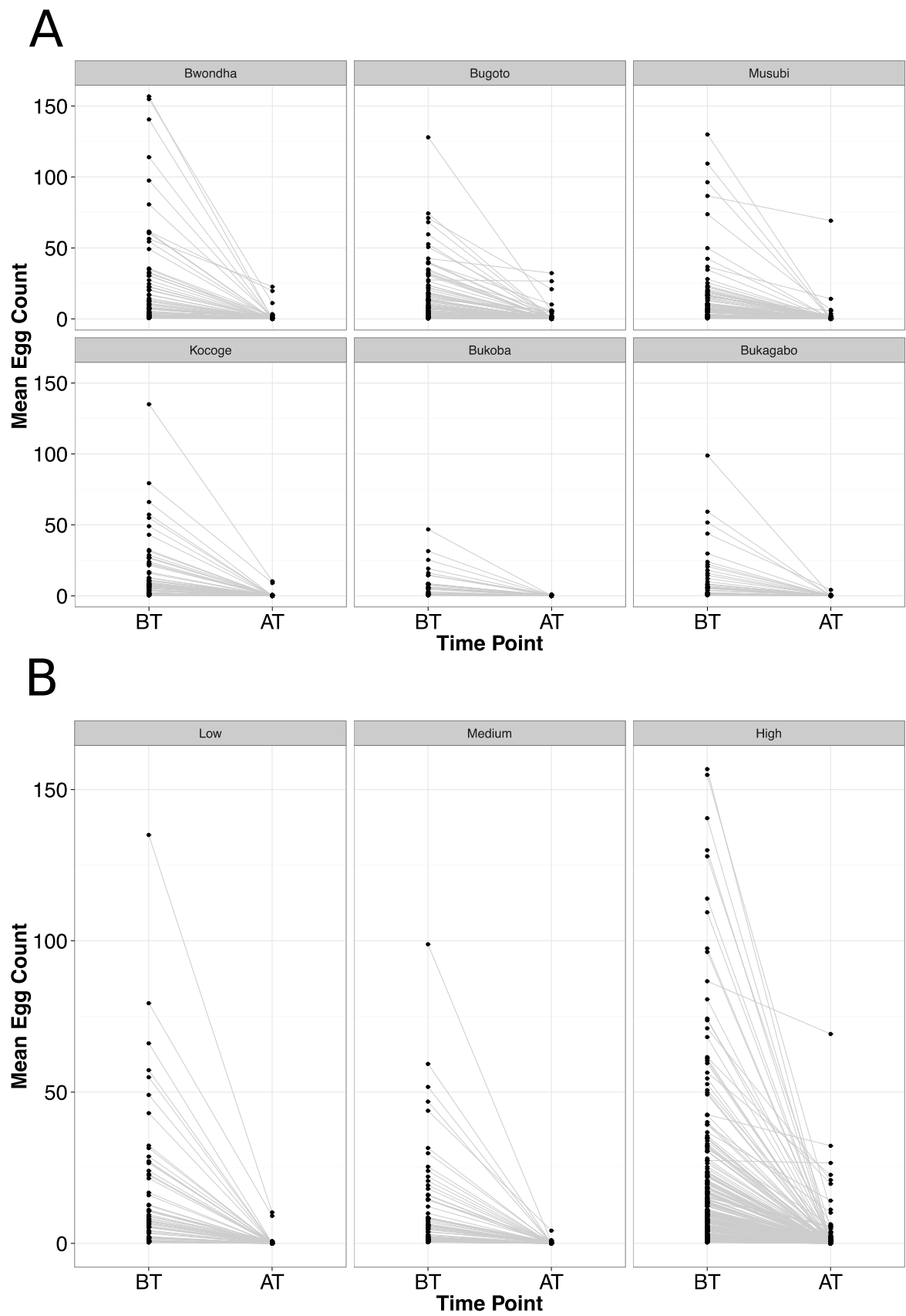
Figure 16. Estimated egg reduction rates (ERRs) of *Schistosoma mansoni* from 414 children in eastern Uganda treated with praziquantel.

Individual estimates are stratified by school in panel A and by mass drug administration (MDA) exposure category in panel B. Black points indicate the mean of the posterior distribution and the grey bars indicate 95% Bayesian credible intervals (BCIs). The hashed lines correspond to the nominal reference threshold of 90% suggested by the World Health Organization. Note that the BCIs have been truncated at 1 and 99%. Category of MDA reflects the previous number of rounds of treatment with praziquantel the schools have received (high is 8 or 9; medium is 5; low is 1).

Figure 17. Estimated egg counts of *Schistosoma mansoni* infections in 414 children in eastern Uganda before (BT) and after (AT) treatment with praziquantel.

Individual estimates are stratified by school in panel A and by mass drug administration (MDA) exposure category in panel B. Category of MDA reflects the previous number of rounds of treatment with praziquantel the schools have received (high is 8 or 9; medium is 5; low is 1).

Figure 17. Estimated egg counts of *Schistosoma mansoni* infections in 414 children in eastern Uganda before (BT) and after (AT) treatment with praziquantel.



Variation in the efficacy of praziquantel among schools

The posterior distributions of the fixed and random effects coefficients were marginalized (averaged) over to yield average ERRs by schools and MDA exposure categories. Average ERRs are generally high and $\geq 90\%$, however schools within the high MDA exposure category show a significantly lower ERR (91.49%; 95% BCI 88.23%, 93.64%) compared with both the medium (98.04%; 95% BCI 96.13%, 99.08%) and low (97.81%; 95% BCI 95.5%, 98.96%) categories, having adjusted for other covariate fixed effects (Table 5). For comparison with the model-derived estimates and for comparability with previous studies, sample estimated ERRs were also calculated. The sample estimated ERRs follow the same pattern as the model-derived ERR estimates, with more intensively treated schools showing lower ERRs, and fall within the corresponding 95% BCI (Table 5). The posterior distributions of average ERRs by school and MDA exposure category are shown in Figure 18. The distribution associated with the high MDA exposure category is separated from the distribution of the medium and low categories, in accordance with a lower average efficacy of praziquantel for children in schools with a higher past exposure to MDA.

Sensitivity of ERR estimates to number of readings taken

The sensitivity of the results to the number of readings per individual were assessed by sub-sampling from the original dataset 1-6 readings before and 1-6 readings after treatment and using the mixed effects model to calculate the estimated ERR for each MDA exposure category. Note that the readings refer to sequential time points (rather than random resampling) so that, for instance, a child may be absent for readings 1 and 2 after treatment but present for readings 3-6. This gives an indication of what the results would show had fewer egg counts been taken and/or sampling had taken place on fewer days. The results are presented in Table 6. The estimated ERR calculated using all readings (up to 6 Kato-Katz slides before and after treatment) as reported in the main text is also included for comparison. In all instances the ERR observed in the High MDA exposure schools remains the lowest. However, a significant difference (i.e. non-overlapping 95% credible intervals) between MDA exposure groups is only observed with 5 or 6 readings. This result indicates that at least 3 days of Kato-Katz readings before and after treatment (with 2 slides per day) was necessary to detect the difference in ERR between MDA exposure categories.

Table 5. Estimated mean egg reduction rates (ERRs) of *Schistosoma mansoni* in 414 children from six primary schools in eastern Uganda following treatment with praziquantel.

			ERR ^a (%)					ERR (%)	
MDA exposure category	<i>n</i>	Cleared (%)	Sample (95% CI)	Model (95% BCI)	School	<i>n</i>	Cleared (%)	Sample (95% CI)	Model (95% BCI)
High (8 or 9 rounds)	268	151 (56.34)	92.10 (87.42, 95.71)	91.49 (88.23, 93.64)	Bwondha	68	40 (58.82)	94.27 (88.13, 98.26)	94.19 (90.65, 96.34)
					Bugoto	103	62 (60.19)	89.63 (77.65, 96.65)	91.20 (85.40, 94.50)
					Musubi	97	49 (50.51)	91.98 (82.35, 97.69)	89.90 (83.53-93.49)
Medium (5 rounds)^b	82	68 (82.93)	98.71 (97.50, 99.53)	98.04 (96.13, 99.08)	Bukoba	42	37 (88.10)	98.98 (97.55,99.85)	98.50 (96.26-99.52)
					Bukagabo	40	31 (77.50)	98.58 (96.50, 99.64)	97.56 (94.59-99.01)
Low (1 round)	64	45 (70.31)	98.87 (98.07, 99.60)	97.81 (95.51, 98.96)	Kocoge	64	45 (70.31)	98.87 (98.07, 99.60)	97.81 (95.51-98.96)

^a Egg Reduction Rate ^b Estimated from district-level MDA coverage from Ministry of Health

Table 6. Estimated mean egg reduction rates (ERRs) of *Schistosoma mansoni* after resampling by Kato-Katz readings.

Cumulative readings (before and after treatment)	1	2	3	4	5	6
Total egg counts	<i>n</i>= 678	<i>n</i>= 1356	<i>n</i>= 2130	<i>n</i>= 2904	<i>n</i>= 3503	<i>n</i>= 4102
MDA exposure category	Mean Egg Reduction Rate (%; 95% BCI)					
High (8 or 9 rounds)	95.97 (87.31, 98.87)	89.75 (80.43, 94.16)	86.31 (72.44, 92.39)	86.89 (77.20, 91.78)	89.64 (83.80, 92.92)	91.49 (88.23, 93.64)
Medium (5 rounds)^a	98.28 (92.61, 99.76)	94.77 (81.67, 98.60)	97.17 (93.70, 98.81)	97.76 (95.54, 98.94)	97.95 (95.84, 99.04)	98.04 (96.13, 99.08)
Low (1 round)	99.13 (95.74, 99.95)	96.36 (84.43, 99.54)	96.48 (88.37, 99.07)	94.34 (81.91, 98.41)	97.47 (94.55, 98.86)	97.81 (95.51, 98.96)

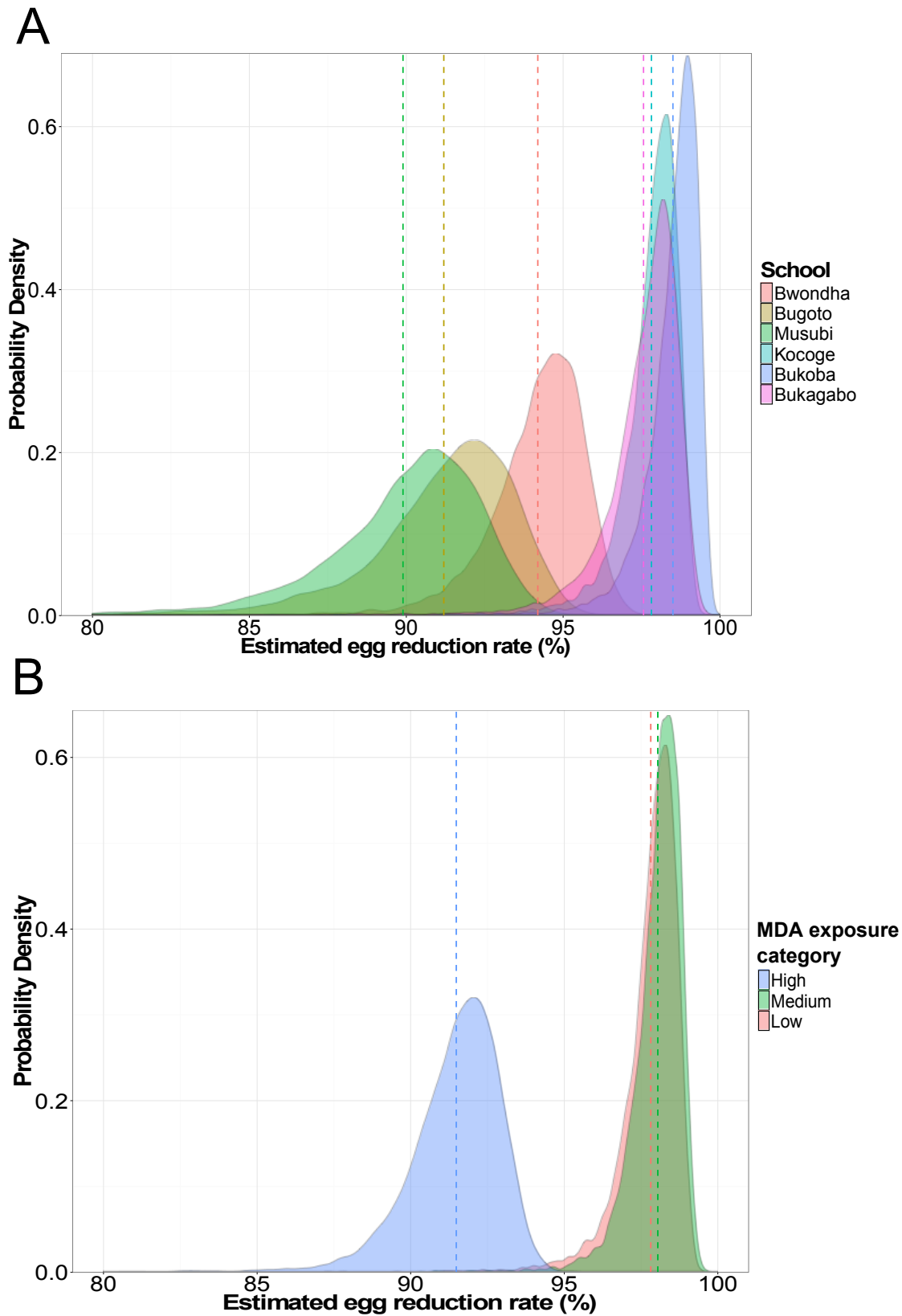
^a Estimated from district-level MDA coverage from Ministry of Health

Table 5. Estimated mean egg reduction rates (ERRs) of *Schistosoma mansoni* in 414 children from six primary schools in eastern Uganda following treatment with praziquantel. Sample estimates are calculated using Equation 2 in Methods, *Sample Egg Reduction Rates* with 95% confidence intervals (CIs) calculated using a block bootstrap approach to account for correlation among egg counts repeatedly measured from the same child. Model estimates and 95% Bayesian credible intervals (BCIs) correspond to the posterior distribution of the average ERR by mass drug administration (MDA) exposure category and by school calculated by marginalizing (averaging) over the appropriate fixed and random effects coefficients of the fitted generalized linear mixed model

Table 6. Estimated mean egg reduction rates (ERRs) of *Schistosoma mansoni* after resampling by Kato-Katz readings. One reading indicates the first reading taken before and after treatment, two readings indicate the first two readings taken before and first two readings taken after treatment etc. Mean egg reduction rates and confidence intervals are presented for MDA exposure categories. The number of samples (n) refers to the total number of egg counts before and after treatment given the cumulative number of readings.

Figure 18. Posterior distribution of average egg reduction rates (ERRs) of *Schistosoma mansoni* in 414 children from 6 primary schools in eastern Uganda treated with praziquantel. The average ERRs are stratified by school in panel A and by mass drug administration (MDA) exposure category in panel B. The posteriors were constructed by marginalizing over the fixed and random effects coefficients estimated from the generalized linear mixed model. The coloured hashed lines indicate the mean of the posterior. Category of MDA reflects the previous number of rounds with praziquantel the schools have received (high is 8 or 9; medium is 5; low is 1).

Figure 18. Posterior distribution of average egg reduction rates (ERRs) of *Schistosoma mansoni* in 414 children from 6 primary schools in eastern Uganda treated with praziquantel.



Discussion

I tested the impact of a school's past exposure to MDA with praziquantel on present treatment outcomes for *S. mansoni* and demonstrated, through the use of an appropriate statistical modeling technique, that treatment efficacy, as quantified by the ERR, is lower in schools that have been treated in eight or nine previous MDA rounds compared with other schools from the same or nearby districts that have undergone five or fewer rounds of MDA with praziquantel.

From the baseline data collected from the schools (Table 3), I found a high prevalence (>84%) and infection intensity (>400 EPG) in Bugoto, Bwondha and Musubi, the schools that have undergone the highest number of previous MDA rounds. It is surprising that *S. mansoni* infection remains entrenched within these foci despite up to nine rounds of MDA. I concur with Kabatereine *et al.* (2003) that persistence of the parasite population is driven by continued re-infection from transmission hotspots in Lake Victoria [77]. It is concerning for efforts to halt transmission of *S. mansoni* through MDA alone that areas with high infection intensity are so resilient to repeated treatments.

The weight of children was found to be associated with estimated ERRs. The dose of praziquantel is weight-standardized at 40mg/kg and hence one would not expect differences in drug efficacy among children of different weight. The lower ERR in lighter children suggests that they are not receiving a sufficient dose, perhaps indicating that there is a minimum therapeutic dose that lighter children are failing to receive. There could also be an age-dependent effect (as weight and age are closely correlated) reflecting synergism between the development of protective immune responses and the antischistosomal activity of praziquantel [206].

A recent meta-analysis of praziquantel efficacy reported an average ERR of 89.1% (95% CI 83.3%, 94.2 %) among treated *S. mansoni* infections in schoolchildren with a median follow up time of 5 weeks [207]. The comparable estimates of the ERR in this study are the average school-level estimates, including those for different MDA exposure categories. None of the estimates extend below the lower CI of the meta-review estimates, suggesting that praziquantel remains broadly efficacious in all the study sites (although heterogeneity in study design will naturally widen CIs estimates from a meta-analysis compared to estimates obtained from a single study). However, of concern is the significant difference in estimated ERRs observed between schools that have

received more previous rounds of MDA compared to those that have received fewer rounds (Table 5, Figure 18).

Variation in anthelmintic efficacy is driven by parasite and host-specific factors. Without long-term longitudinal data (i.e. on the scale of years, over multiple rounds of MDA), or baseline data from before MDA commenced, it cannot be determined whether heterogeneities among ERRs, at either the individual or school levels, are static or changing temporally. A systematic decline in ERRs could be indicative of increasing parasite tolerance or emerging resistance to praziquantel [23, 135]. Intervention programmes using MDA should incorporate regular assessment of drug efficacy into their monitoring and evaluation protocols.

A recent analysis of praziquantel efficacy against *S. mansoni* found variation among ERRs from individual patients to be greater in Uganda than in other countries [84]. *Schistosoma mansoni* parasites in Uganda have been observed to be more genetically diverse than elsewhere [20], which could be driving more variable treatment outcomes. Host factors such as genetic variation or differences in nutritional status are unlikely to be sufficient to explain the variation in ERRs because of the spatial proximity and similar patterns of subsistence in the schools surveyed (Figure 13). It is also unlikely that different rates of reinfection are driving substantive differences among patient ERRs. This is because the time to follow up (25-27 days) is short and the statistical model is designed to control for different rates of reinfection by modeling the intensity of infection after treatment *relative* to the intensity before treatment such that each individual acts as its own control.

We are left with the possible explanation that reduced praziquantel efficacy is caused by drug-tolerant genetic variants within the *S. mansoni* population in highly treated areas. Genetic changes associated with drug resistance have been associated with drug-induced selection pressures in human and veterinary helminths [113, 208-210]. Ultimately, drug-induced resistance can only be confirmed by linking changes in resistance allele frequencies to phenotypic indicators of drug tolerance or reduced susceptibility. As neither the precise mode of action of praziquantel nor the genetic basis of resistance in schistosomes are completely understood [211], the powerful approaches offered by NGS will be essential to scan for selection across whole-genomes [132,

140]. Such approaches would help to establish if parasite genetic variation could be driving the degree of variation in praziquantel efficacy observed in this study.

Ethical Statement

Approvals were granted by the Uganda National Council of Science and Technology (MoU sections 1.4, 1.5, 1.6) and the Imperial College Research Ethics Committee (EC NO: 03.36. R&D No: 03/SB/033E). Verbal assent was given by every child before inclusion into this study and at school committee meetings comprising of parents, teachers, and community leaders before the onset of the study. Written consent for the children to participate in the study was attained from each head teacher. Participation was voluntary and children could withdraw or be withdrawn from the study at any time. All enrolled children were treated with 40 mg/kg praziquantel and 400 mg albendazole after examination at baseline and any children who remained positive at follow up were retreated.

Chapter 4

Whole-genome sequencing and population genomics of *Schistosoma mansoni* miracidia from primary school children in Uganda

Abstract

In the previous Chapter I demonstrated that the efficacy of praziquantel against *Schistosoma mansoni* was associated with the number of rounds of mass drug administration in primary schools in eastern Uganda. It is therefore of considerable interest to interrogate parasite genomes from patients with variable clearance phenotypes to understand the population structure, the extent to which positive selection is acting on the *S. mansoni* genome and whether variants can be associated with the host parasite clearance phenotype.

Whole genome sequences were generated from 213 miracidia from 35 children from four schools in eastern Uganda. After calling variants and applying filters, we obtained a dataset of 5.31 million SNP variants from 190 miracidia. Examining the population structure with principal component analysis, F_{ST} and IBD revealed a largely unstructured parasite population at the level of individual hosts and schools. Genome-wide scans for signatures of positive selection using F_{ST} and a haplotype based test (iHS) revealed a 256kb region at the beginning of chromosome 2 to show the most extreme values in both tests. Association testing of the SNPs in the highest 1% of regions under selection against the host clearance phenotype revealed 4 SNPs to be significantly associated after correcting for multiple testing, 3 of which were located in the 256 kb region in chromosome 2. Four genes are predicted to be within this region, two of which are annotated as calcium/ sodium channels. While the precise mode of action of praziquantel is unknown, there is strong experimental evidence linking the action of the drug on *S. mansoni* to calcium signalling. Within the exons of the calcium channels, I uncovered six putative causal variants (non-synonymous mutations and indels) one of which was significantly associated with the host clearance phenotype.

By identifying a region of the *S. mansoni* genome under positive selection containing calcium channels and significantly associated with the host clearance phenotype, this study provides convincing evidence that repeated MDA with praziquantel is driving selection in parasite genomes towards a phenotype that inhibits the action of praziquantel. These data provide the rationale for

future studies to validate the function of the genes in this region. If these genes are in fact shown to regulate uptake of praziquantel, we may finally begin to understand the drug's mechanism of action, which has remained unknown since praziquantel was synthesised over 40 years ago.

Introduction

In the previous Chapter, I demonstrated that children infected with *Schistosoma mansoni* in intensively treated schools in eastern Uganda had a reduced response to the drug praziquantel (as measured by the egg reduction rate (ERR)) compared with children in less intensively treated schools [92]. This significant difference in parasite clearance was observed despite the close spatial proximity of the schools suggesting that the differences observed can be explained, at least partly, by variation in the parasite population as opposed to the host population which we assume is largely uniform across the sampled regions [101]. It is therefore of considerable interest to understand the population genetic structure of the parasites within the sample of infected individuals in Chapter 3 and to determine if there is a genetic basis in *S. mansoni* that accounts for the differences in anthelmintic efficacy we have observed at the host level.

Genome-wide analyses of genetic variation (i.e. association studies, scans of selection or genetic mapping) are increasingly being used to uncover the genetic basis for traits of interest such as virulence and drug resistance in parasites and pathogens [133, 137, 212]. Whole-genome sequencing of *S. mansoni* parasites directly from clinical samples is a powerful method to understand the dynamics shaping the population [7] and allows us to utilise unbiased ('hypothesis free') methods with which to identify interesting regions of the genome for future functional validation [138, 140, 213]. In general, these methods exploit the linkage disequilibrium of regions under selection. As selection acts to increase the frequency of variants more quickly than linked regions can be broken down by recombination, in the short term selection will act on the entire haplotype surrounding a *de novo* variant of high fitness, rather than affecting just the variant position itself. Within a population under selection therefore, the *de novo* variant will appear on a longer, less variable haplotype compared with a region evolving neutrally [149, 213]. This case is complicated in the case of a 'soft sweep', whereby selection acts on standing variation (i.e. variation existing before the selective pressure appeared), rather than a *de novo* mutation, as most methods lack power to detect soft sweeps. To date, no genome-wide scans of selection and genetic

association studies have been performed on schistosome parasites obtained directly from infected individuals, and therefore, the data presented in this Chapter offers novel insight into the genetics of praziquantel response by the parasite. Although no specific loci in *S. mansoni* have previously been associated with praziquantel response, it has been demonstrated that treatment with praziquantel results in a rapid Ca^{2+} influx leading to sustained Ca^{2+} dependent muscular contraction [214]. Further, treating *S. mansoni* with chemicals that inhibit the action of calcium channels has been found to reduce the sensitivity of the parasite to praziquantel [78, 79, 81].

Based on the evidence for reduced efficacy of praziquantel in intensively treated schools and our limited knowledge of praziquantel's mode of action, I hypothesise that the poor responder phenotype of *S. mansoni* to praziquantel is associated with: i) gene(s) that show signals of positive selection in the population as a result of the repeated rounds of mass drug administration (MDA) with praziquantel; and ii) genetic variation within these gene(s) that is associated with the clearance phenotype of the host, measured either as the quantitative ERR estimated from the mixed-effects model in Chapter 3, and/or, as a categorical host trait (i.e., (a) good clearer before treatment; (b) poor clearer before treatment; (c) poor clearer after treatment). By identifying genes and genetic variation using these criteria, together with prior knowledge of the action of praziquantel on schistosomes (i.e., strong candidate genes would be expressed in the adult worm as this is the life stage in the *S. mansoni* lifecycle against which the drug acts and could be involved in calcium signalling), it should be possible to prioritise candidate genes for further validation towards understanding the mechanism of praziquantel action and poor response.

In addition to documenting genetic variation under selection in *S. mansoni*, whole genome re-sequencing of parasites from individuals in Uganda allows us to understand the population structure of the parasite from a natural foci of infection. Previous studies of the population structure of *S. mansoni* based on microsatellites or barcoding regions have shown an absence of population structure within countries [101, 112, 215], which is perhaps surprising given the portioning of parasites into host 'demes' that may impede gene flow [216]. The genetic structure of the parasite population can determine how positive selection acts upon a population and how variants conferring higher fitness can spread; documenting the spatial genetic structure of *S. mansoni* populations using whole genome SNP data would therefore add greatly to our understating of how genetic variance is partitioned in these populations.

As discussed in the Introduction Chapter, previous studies examining genetic markers of *S. mansoni* parasites have differed on whether a reduction in genetic diversity was observed following treatment [113, 114, 217]. Individuals infected with *S. mansoni* may not experience complete parasite clearance after treatment with praziquantel, as measured by eggs in stool (as discussed in Chapter 3) or through hatching miracidia from stool [86, 113], though it is not always clear whether this represents parasites surviving treatment or re-infection post-treatment. Therefore, I aim to address this question by sequencing parasites from before and after treatment in the same individuals to determine: i) if a reduction in genetic diversity can be observed after a single round of treatment; and ii) if parasites taken from before and after treatment within individual human hosts are more closely related than to the parasite population as a whole, which could indicate persistence of infection rather than re-infection.

While there is great potential for applying whole genome sequencing to *S. mansoni* parasites from natural populations, there are a number of difficulties in generating and analysing the necessary sequence data. Ideally, adult parasites would be sequenced as these represent the target of the drug praziquantel, however, they cannot be obtained from living humans [218]. Therefore, sequencing of parasites directly from individuals requires isolating miracidia (offspring of adult worms) from parasite eggs in stool samples, which represents an indirect sampling of genetic variation associated with response as the adult worms have undergone sexual recombination. Secondly, miracidia contain fewer than 100 cells and therefore only small amounts of DNA are available to be sequenced; this limits the extent to which high quality whole-genome sequence data can be generated. Furthermore, contaminating DNA also present in faeces, including host and bacterial genetic material, may overwhelm the parasite material when DNA extracted from stool samples is PCR amplified and/or sequenced without specific enrichment for *S. mansoni* DNA. Another limitation is that DNA sequencing libraries prepared from low quality or low quantity DNA typically requires substantial PCR amplification and often result in 'low complexity' libraries that contain a high proportion of duplicate reads. Duplicate reads, which are technical artefacts that result from an over-enrichment of a limited number library reads can lead to uneven depth along the genome, which often confounds variant calling as sequencing errors may become enriched and appear as true variants when called against a reference genome. This in turn can bias downstream analysis [219]. Finally, the quality of the reference genome to which the

sequenced reads are aligned is also a major factor in the quality of downstream analysis. Many of these challenges are magnified in *S. mansoni*, relative to most other pathogens, by the relatively large size of the genome (369 mb in version 6), ~80 times larger than bacteria such as *E. coli*, ~16 times larger than *P. falciparum* and ~4 times larger than nematodes such as the parasitic *Brugia malayi* or the free-living *C. elegans*. This therefore increases the cost and complexity of analysis and sequencing.

In this Chapter, I describe how I developed a novel wet-laboratory protocol aimed to generate, and subsequently analyse, viable whole-genome sequence data from individual miracidia isolated from human stool samples in the field in Uganda. My specific aims were to elucidate: i) the population structure of natural *S. mansoni* populations under contrasting mass drug administration selective pressures at the level of whole genome variation; ii) the genome-wide variability in miracidia collected from before and after treatment from individuals showing suboptimal clearance; iii) regions of the genome showing signals of positive selection; and iv) whether *S. mansoni* variants under selection are associated with the praziquantel efficacy phenotype of the host.

Methods and Materials

Sample collection in the field

During the fieldwork in eastern Uganda from May-June 2014 described in Chapter 3, the field team and myself hatched and stored over 10,000 miracidia from children in the five schools in Mayuge District (Bugoto, Bwondha, Musubi, Bukoba and Bukagabo) and one school in Tororo District (Kocoge) (see Figure 13, Chapter 3 for map of Uganda). Miracidia were collected from a subset of the 414 children included in the drug efficacy analysis from Chapter 3, as time and logistical constraints prevented us from collecting parasites from all the enrolled children. We collected miracidia from children before and 25-27 days following treatment with praziquantel in the case of incomplete clearance. This time to follow up was chosen to give enough time for the drugs to act, while minimising the risk of re-infection [23]. Given that only a small proportion of individuals showed incomplete clearance and we had not sampled from all individuals before treatment, we were only able to collect miracidia from both time points from a small number of individuals within incomplete parasite clearance (see below). To collect the miracidia, stool samples were

filtered through an adapted Pitchford and Visser funnel and washed with bottled water to dilute occluding faecal debris [101, 220]. To hatch the miracidia, the final elute containing schistosome eggs was poured into petri-dishes and left overnight, before being exposed to sunlight for hatching the following morning. A minimum of 20 miracidia were collected from individuals diagnosed as positive for *S. mansoni* with Kato-Katz at each time point to capture within-host variability, however, due to cost constraints we only sequenced a maximum of 7 miracidia per patient, per time point. The miracidia were isolated with a 20µl pipette and light microscope (Figure 19). Parasites were pipetted from the petri-dish and placed in a fresh petri-dish of nuclease-free water to dilute bacterial contaminants, before being pipetted again and spotted onto a Whatman® FTA® card, which lyses cells and organelles and immobilises and preserves nucleic acids [221, 222]. The FTA® cards can be kept at room temperature and are light and transportable, meaning they provide a convenient method of storing DNA in the field as they remove the need for refrigeration or freezing. Miracidia were isolated from the stool of 202 individuals before treatment, 54 individuals after treatment and an overlapping subset of 21 individuals both before and after treatment. The number of miracidia collected ranged from 20-100 per child, per time point.

Protocol development for DNA extraction / amplification

It was necessary to develop a protocol to extract and amplify DNA from miracidia on FTA® cards for whole-genome sequencing as this had never previously been attempted. Prior to large-scale sequencing, the protocol was validated by sequencing samples on the MiSeq platform (Illumina) at low depth. Although the sequence data from these runs were not usable in downstream analysis due to the very low depth of coverage (<1x), the data provided a valuable indicator of the degree of contamination in each sample together with the complexity of the library when the reads were mapped to the *S. mansoni* v5.2 reference genome [124].

Initially, I used a protocol to amplify the miracidium DNA directly on the FTA® card punch. A single miracidium was removed from FTA® card using a Harris micro punch and washed in TE and FTA® buffers, after which the buffers were aspirated. Reagents from the REPLI-g® Mini kit (Qiagen) were added directly to the punch following manufacturers instructions. This involved a denaturation step to remove any bound proteins from the DNA, then polymerase was added and the sample incubated at 30°C for 9 hours. Finally, clean up of the amplified product was performed using magnetic AMPure® XP beads (Agencourt®; following manufacturer's protocol) that bind

DNA and allow any remaining reagents to be aspirated. However, this method failed to amplify sufficient DNA to be detectable by Qubit® dsDNA HS (High Sensitivity) Assay Kits (Thermo Fisher Science), indicating that there was less than 0.5 ng/ml DNA present (detection limit of the kit).

A



B



C



Figure 19. Technician Andrina Nankasi collecting miracidia during field studies in eastern Uganda.

Miracidia were pipetted into nuclease free water with a 20 μ l pipette and visualised under a light microscope (Panel A), before being re-collected (Panel B) and spotted onto a Whatman FTA® card (Panel C). Photos taken by Thomas Crellen.

To better optimise the protocol for DNA extraction and amplification, I performed a number of experiments using miracidia from Tanzania that had been collected in 2015 using a number of different methods: miracidia were washed in either: i) nuclease-free water; ii) salt buffer; or iii) antibiotics (Pen Strep (Thermo Fisher Scientific) and Gentamicin (Thermo Fisher Scientific)). These miracidia would subsequently be stored in either 10µl DNA/RNA Shield™ (Zymo), a buffer that lyses cells and preserves DNA ($n=7$) or on FTA® cards as before ($n=25$). These samples were brought to WTSI, where they were extracted with the use of the DNeasy Blood & Tissue Kit (Qiagen), which removes any proteases and membranes before purifying the DNA through a spin column. The eluted DNA was then amplified using one of three methods: 1) REPLI-g® Mini Kit ($n=14$); 2) illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences) ($n=7$); or 3) the DNA elute was first concentrated by drying down on a Thermo Scientific™ Savant™ SPD131DDA SpeedVac™ Concentrator and re-suspending in nuclease free water for 1 hour prior to amplifying with GenomiPhi V2 DNA Amplification Kit (this method is abbreviated to “concentrated GenomiPhi”; $n=11$). For these experiments, the punches of miracidia were pooled into lots of 10 for extraction and amplification to maximise the chance of extracting *S. mansoni* DNA. Of the 32 pools of samples amplified, 14 were miracidia from Tanzania and 18 were miracidia from Uganda (as described earlier). After amplification, the 32 pools were sent to the libraries team at WTSI and sequenced on an Illumina MiSeq with 150bp paired end reads; libraries were prepared with MiSeq Reagent Kits v3 (Illumina). Across all collection methods, the proportion of reads mapping to *S. mansoni* reference genome were within the range 1.0%-93.5% (median; 33.8%). A linear model was constructed and fitted in R version 3.2.2, relating the proportion of mapped reads to the amplification method (REPLI-g®, GenomiPhi or concentrated GenomiPhi), the country collected (Tanzania or Uganda), the storage method (FTA® card or DNA/RNA Shield™), and the treatment of the samples on collection (1x wash, 3x wash, salt buffer or antibiotics). Using a significant threshold of $p < 0.05$, the only covariates to have a significant effect on the outcome variable was the amplification kit. When compared against REPLI-g®, the GenomiPhi (coefficient = 39.9, $p = 0.01$) and concentrated GenomiPhi (coefficient = 53.6, $p = 0.0002$) methods resulted in a higher proportion of *S. mansoni* reads mapping (see Figure 2). This was a positive outcome as it suggested that, rather than needing to collect additional field samples with a novel collection method, by using a different amplification kit I would have more success in generating DNA from miracidia that would allow whole-genome sequencing at higher depth to be viable.

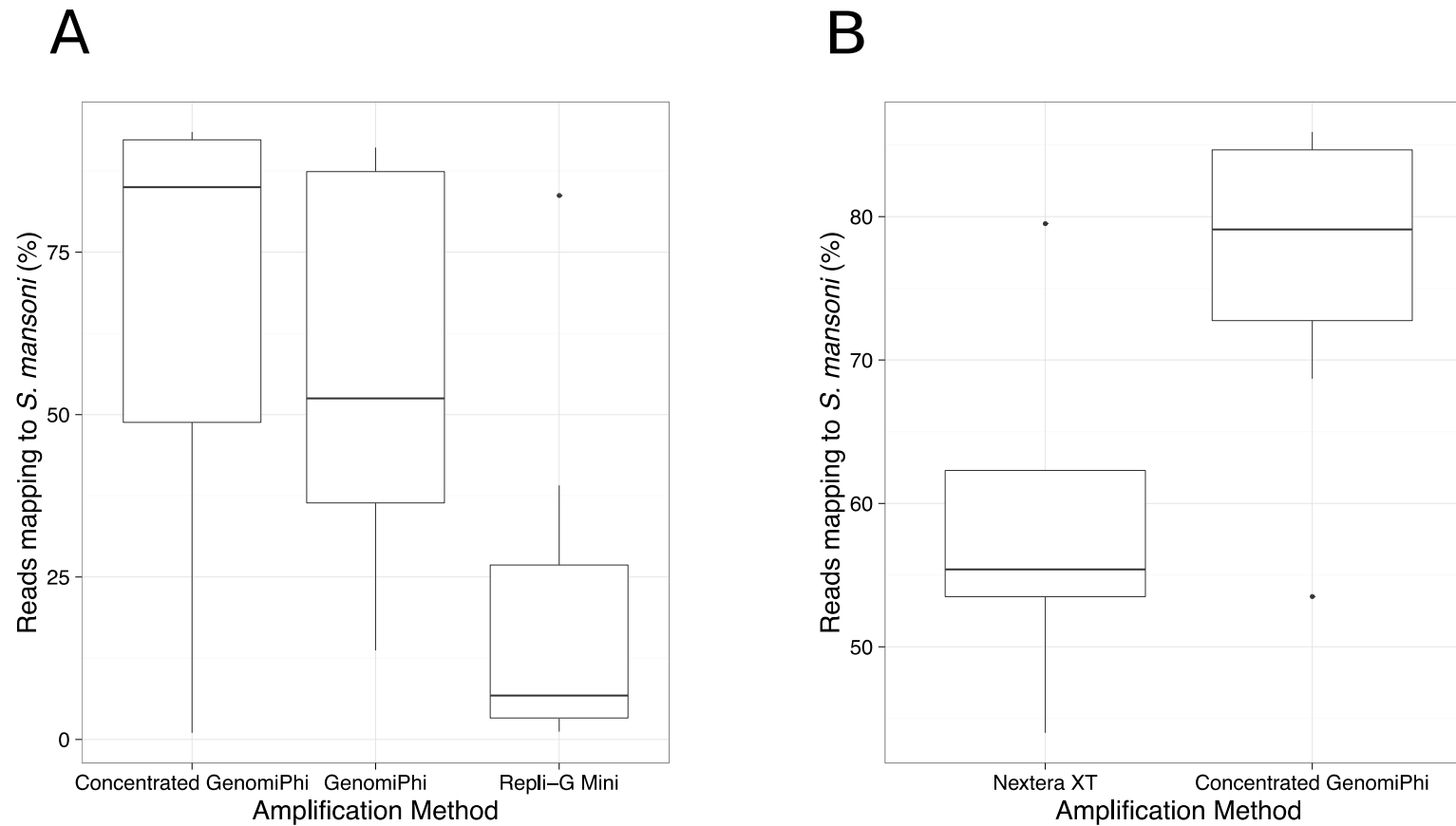


Figure 20. Boxplots comparing percentage of reads mapping to *S. mansoni* v5.2 reference genome by amplification method. Panel A shows a comparison between the concentrated GenomiPhi, GenomiPhi and REPLI-g® methods, when miracidia are pooled in lots of 10 for extraction, amplification and sequencing. Panel B shows boxplots comparing percentage of reads mapping to *S. mansoni* version 5.2 reference genome by method (concentrated Genomiphi WGA and Nextera XT) from single miracidia.

I proceeded to examine whether it would be possible to attain sequence data from individually amplified miracidia, rather than from multiple samples in a pool as before. Sequencing individual miracidia is preferable, as when amplifying multiple samples the resulting sequence data may not give representative allele frequencies if the quantity of DNA extracted or the efficiency of subsequent amplification varies between miracidia. The concentrated GenomiPhi method was therefore applied to 20 miracidia from FTA® cards, of which 16 were successfully amplified (>0.5 ng/ml). An additional QC step was also introduced at this stage, which involved PCR amplification of the mitochondrial cytochrome oxidase 1 gene (*cox1*) to test for the presence of *S. mansoni* in the sample, as during the picking of miracidia in the field it is possible for technicians to miss a miracidium so that only water is spotted onto the FTA® card. A 25µl PCR reaction was prepared with 1ul of product (DNA), 10µl GoTaq® G2 Green Master Mix (x2), 1µl Forward Primer 12.5uM (100uM stock diluted 1:7), 1µl Reverse Primer 12.5uM (100uM stock diluted 1:7), 7µl nuclease-free H₂O. The PCR primers for *cox1* have been published previously [102] and the thermocycler (Eppendorf™ Mastercycler™) conditions were as follows: A single incubation at 95°C for 5 minutes, followed by 30 cycles of 95°C for 30 seconds, 63°C for 30 seconds, 72°C for 30 seconds, and a final incubation at 72°C for 5 minutes.

Of the 16 samples with amplified DNA, 10 were PCR positive for *S. mansoni cox1* as visualised on a 1.5% agarose gel. When sequenced at low-depth on a MiSeq (Illumina), the range of reads from the 10 libraries mapping to *S. mansoni* were 53.5% - 85.9% (median = 77.5%), and generally contained a low number of duplicate reads (<3%) with the exception of one sample. Analysing the distribution of reads mapping along chromosomes showed that reads were sequenced evenly across the genome and were not clustered in a specific region. These results were promising and suggested that this method could be expanded beyond the pilot to generating high quality sequence data from individual miracidia on the higher throughput HiSeq X10 (Illumina) sequencing platform.

The viability of an alternative amplification method, the Nextera XT DNA Library Preparation Kit (Illumina), was also examined. In this method a transposable element is used to fragment and tag DNA with adaptors in a single step [223]. Adaptor bound libraries are then amplified using 8 cycles of PCR. Library concentration was visualised on an 2100 Bioanalyser Instrument (Agilent) to determine if Nextera libraries had successfully bound to DNA and were then tested for the

presence of *S. mansoni* with a PCR amplification of the *cox1* gene using the conditions described above. Starting with 10 individual miracidia punches (extracted with the DNeasy Blood and Tissue Kit prior to the Nextera protocol), in five samples the Bioanalyser (Agilent) detected Nextera libraries, and these samples were subsequently found to contain *S. mansoni* DNA, as detected by PCR amplification of the *cox1* gene. After sequencing at low depth on the MiSeq, the percentage of reads mapping from the five samples to the *S. mansoni* reference genome ranged from 44.0-79.5% (median 55.4%). Comparing the WGA ‘concentrated GenomiPhi’ method against Nextera, samples amplified with WGA prior to sequencing on the MiSeq had a higher proportion of reads mapping to *S. mansoni* (range 53.5% - 85.9%, median 77.5%; Figure 20). Therefore on the basis of these results, it appeared that the concentrated GenomiPhi method was better suited to generate quality sequence data.

Sequencing at depth

Initially three 96 well plates were prepared using the concentrated GenomiPhi method described above, however, quality control (QC) revealed very low DNA yields even after multiple rounds of PCR amplification of adapters. To determine the cause of these low yields, I observed a strong plate-dependent effect as the DNA yields varied by plate location suggesting that the ‘concentrating step’ in the heated vacuum had damaged the DNA prior to amplification and sequencing. From the first three plates, only 50 samples were deemed to have a high enough DNA yield for sequencing on the Illumina HiSeq X10 platform. Another two plates were prepared, with an amendment made to the protocol; the extracts were only partially dried down, rather than being completely dried down and re-suspended in clean water as before to avoid damage to the DNA. Quality control during library preparation revealed these plates to have higher DNA yields and so samples were divided into two sequencing run depending on DNA/ library concentration. Samples with high DNA/ library concentration were subjected to 8 cycles of PCR, whereas samples with lower yields were amplified using 18 cycles of PCR (see Appendix 1 for full protocol).

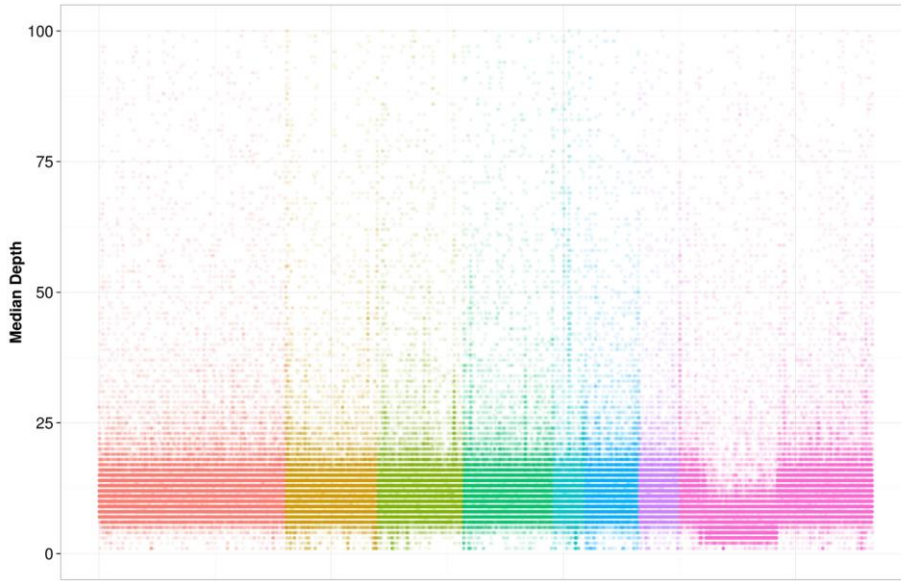
Sample selection, mapping and depth

Miracidia were sampled purposively from individuals with different clearance phenotypes, i.e., those that showed “good” clearance (i.e. no miracidia hatched after treatment) and those that showed “poor” clearance (i.e. it was possible to hatch miracidia from stool samples following treatment), see Table 7 and Figure 23. The ERR for “good” clearing individuals ranged from 99.4-

99.9% (median=99.8%) and for “poor” clearing individuals from 24.2-99.4% (median= 91.1%). In the previous Chapter we defined an acceptable ERR of >90%, therefore it was perhaps surprising that relatively large numbers of miracidia (≥ 20) could be hatched from stool from individuals following treatment that had apparently high ERRs, this suggests that hatching miracidia may be a more sensitive diagnostic than Kato-Katz. From individuals with “poor clearance” we sequenced miracidia from both before and after treatment in the same individuals, so as to compare the genetic variability of miracidia populations from different time points.

In total, 213 miracidia were sequenced on the Illumina HiSeq X10 platform at the Wellcome Trust Sanger Institute using 125bp paired-end reads with the aim of achieving 30x genome coverage. Sequenced reads were mapped to the version 6 *S. mansoni* reference genome using smalt (<http://www.sanger.ac.uk/science/tools/smalt-0>) with options -i 2000 -r 1 -x -y 0.85. These options indicate that reads were mapped to genome positions with a similarity of > 0.85, and in the case that reads map equally well to multiple positions in the reference genome, they will be randomly assigned to one of these positions, thus preventing repetitive regions from having unusually high or low depth. This version of the *S. mansoni* reference was finished in late 2016 with long read (PacBio) sequencing and >90% of the genome had been assigned to chromosomes. It is an improvement on version 5.2 that was used for mapping in Chapter 2. Duplicate reads in the mapped BAM files were removed using Picard v1.92. The median depth of the autosomal chromosomes ranged from 1-38x (interquartile range; 4-14x); see Figure 21 below. While we had aimed for 30x coverage, the lower depth in all but one sample was the result of low quality DNA used to create libraries and subsequently a high proportion of PCR duplicates.

A



B

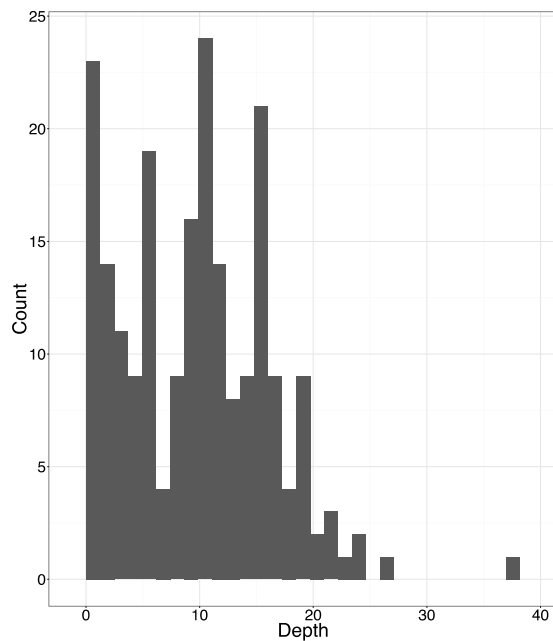


Figure 21. Plot showing sequencing depth (median 11x) for sample Bw9_D along 7 autosomal chromosomes and ZW sex chromosome.

Each point represents the median depth of a 1 kb window and points are coloured by chromosome. The positions of points along the y-axis represent the relative position along each chromosome. The drop in depth in ZW chromosome indicates that this miracidium is female as there is lower coverage over the Z chromosome specific region (Panel A). Histogram of median depth along 7 autosomal chromosomes for 213 miracidia sequenced on HiSeq X10, median depth ranges from 1-38x; IQR: 4-14x (Panel B).

SNP calling QC

Given the lower depth of the BAM files compared with previous work in Chapter 2, where the depth of samples ranged from 23-58x, and considering the present dataset is larger and better represents a population as multiple hosts were sampled within a smaller geographic region, a novel SNP calling pipeline was developed based upon the 'best practises' guidelines produced by GATK v3.6 [224]. Variants were called independently for each sample using HaplotypeCaller then the variants files for each sample (genomic variant call format; gVCF) were merged in batches of 20 before calling SNPs across the cohort with GenotypeGVCF. Indels were then removed leaving only SNPs. Hard filtering of SNPs was performed in two phases: i) filtering based on quality control metrics of the BAM file in GATK; and ii) filtering based on population parameters such as minor allele frequency (MAF), missingness of variants and missingness by sample in PLINK v.3.40 [225]. Filtering in GATK was based on six parameters:

- Quality by Depth (QD): variant confidence divided by depth of non reference samples (filter value ≤ 2.0)
- Fisher Strand (FS): phred-scaled p value of Fisher's Exact Test to detect strand bias (variants observed preferentially on the forward or reverse strand), which indicates false positive SNP calls (filter value ≥ 60)
- Strand odds ratio (SOR): another method estimating strand bias based on the odds ratio (filter value ≥ 3)
- RMS Mapping Quality (MQ): root mean squared mapping quality over all the reads at the site (filter value ≤ 40)
- Mapping Quality Rank Sum (MQRankSum): rank sum test comparing the mapping quality of the reads supporting the reference allele compared against reads supporting the alternative allele (filter value ≤ -10)
- Read Position Rank Sum (RPRS): rank sum test comparing the site position within reads of reference and alternate alleles, as observing alleles only at the end of reads is indicative of errors due to the higher error rate at the end of reads in Illumina sequencing platforms (filter value ≤ -6)

The distributions of each of these parameters were visualised and compared against examples from GATK best practises to optimise the filter cut-off values [224]. These filters resulted in a reduction in the number of SNPs from 17.05 million to 10.61 million. The remaining variants were

then filtered at the population level using PLINK (v3.40) [225]. Three filters were applied: first individuals that had a high proportion of missing variants ($\geq 50\%$) were removed (i.e. the miracidia had insufficient depth to call SNPs at these positions); this reduced the number of miracidia in the population from 213 to 190. Variants with a low MAF ($\leq 1\%$) are likely to contain a high proportion of sequencing errors and so were removed (see MAF distribution in Figure 22). Variants with a high proportion of missingness ($\geq 10\%$) were also removed. These two filters reduced the number of SNPs in the sample from 10.61 to 5.31 million (

“Good” refers to hosts, and parasites from hosts, which cleared completely following treatment with praziquantel. Hosts, and parasites from hosts, which did not fully clear following treatment are classified as “Before” or “After” depending on the time point from which they were sampled.

Table 8). The dataset used for the majority of analyses therefore consisted of 5.31 million SNPs from 190 miracidia.

Additional sequence data from *S. mansoni* adult worms collected in Cameroon, Lake Albert and Lake Victoria (described in Chapter 2) were also used to place the population structure of the Ugandan miracidia in a broader African context as well as to confirm the reliability of the global population structure described in Chapter 2 that used lab-passaged worms [132]. The genomes of these three adult worms were mapped to the V6 *S. mansoni* reference genome (as previously in Chapter 2 they had been mapped to the older V5.2 *S. mansoni* reference) using smalt as described above and SNPs were called for each sample individually using HaplotypeCaller. Variants were then called at the population level for the 213 Ugandan miracidia and the three additional samples using HaplotypeCaller. Filters were applied to these data in PLINK; indels, samples with missingness $> 60\%$ and variants with missingness $> 20\%$ were removed leaving 15.5 million SNPs from 203 individuals.

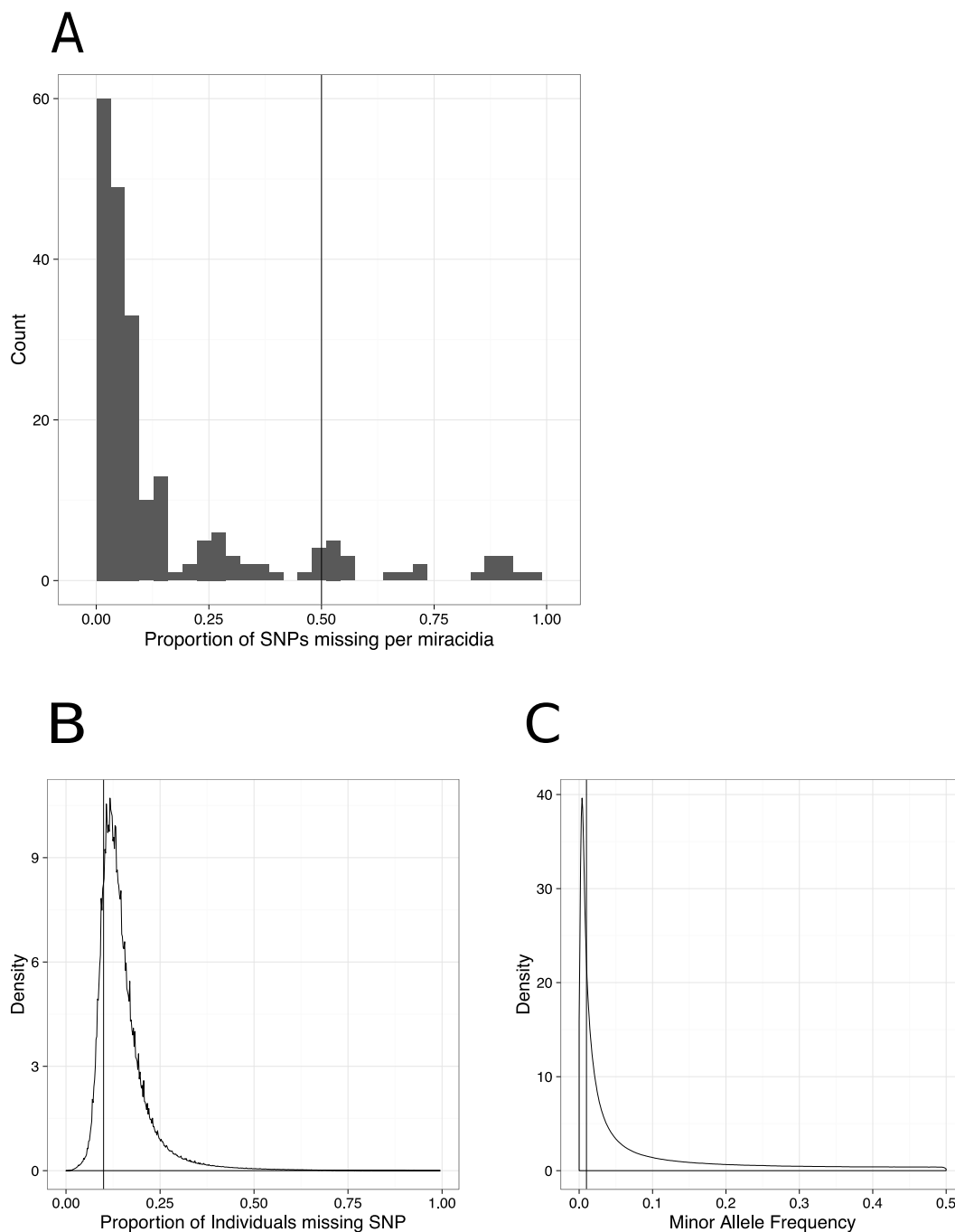


Figure 22. Comparison of sample and genotype missingness, and minor allele frequency using 10.61 million variants and 213 miracidia.

The proportion of missing variants per miracidium is shown in panel A, a cut-off of 50% missing variants was applied removing 23 miracidia is shown. The proportion of missingness by variant is shown in panel B, any variants with missingness >0.1 were filtered, removing 2.61 million variants. The probability density of the minor allele frequency (MAF) is shown in panel C. SNPs with MAF <0.01 were filtered, removing 2.69 million variants.

Table 7. Sampled hosts ($n=35$) and miracidia ($n=190$) mean, range and total by school and time point.

Method	Variants Removed (millions)	Variants Remaining (millions)	Samples (n)
Haplotype Caller (SNPs & Indels)	NA	18.52	213
Remove Indels	1.47	17.05	213
Hard Filtering (GATK)	6.44	10.61	213
Population Filtering (PLINK)	5.30	5.31	190

“Good” refers to hosts, and parasites from hosts, which cleared completely following treatment with praziquantel. Hosts, and parasites from hosts, which did not fully clear following treatment are classified as “Before” or “After” depending on the time point from which they were sampled.

Table 8. Single nucleotide polymorphisms and samples present in the dataset after applying filters from the Genome Analysis Toolkit (GATK) and PLINK.

School	Number of hosts			Number of miracidia per host: median and range			Number of miracidia: total		
	Good	Before	After	Good	Before	After	Good	Before	After
Bugoto	4	7	9	5 (2-7)	4 (1-6)	3 (1-5)	19	25	28
Bwondha	6	4	3	5.5 (1-7)	4 (4-5)	2 (2-4)	30	17	8
Musubi	6	3	3	3.5 (2-5)	5 (1-6)	4 (3-6)	20	12	13
Kocoge	3	0	0	6 (5-7)	0	0	18	0	0
Total	19	14	15	5 (1-7)	3 (1-6)	4 (1-6)	87	54	49

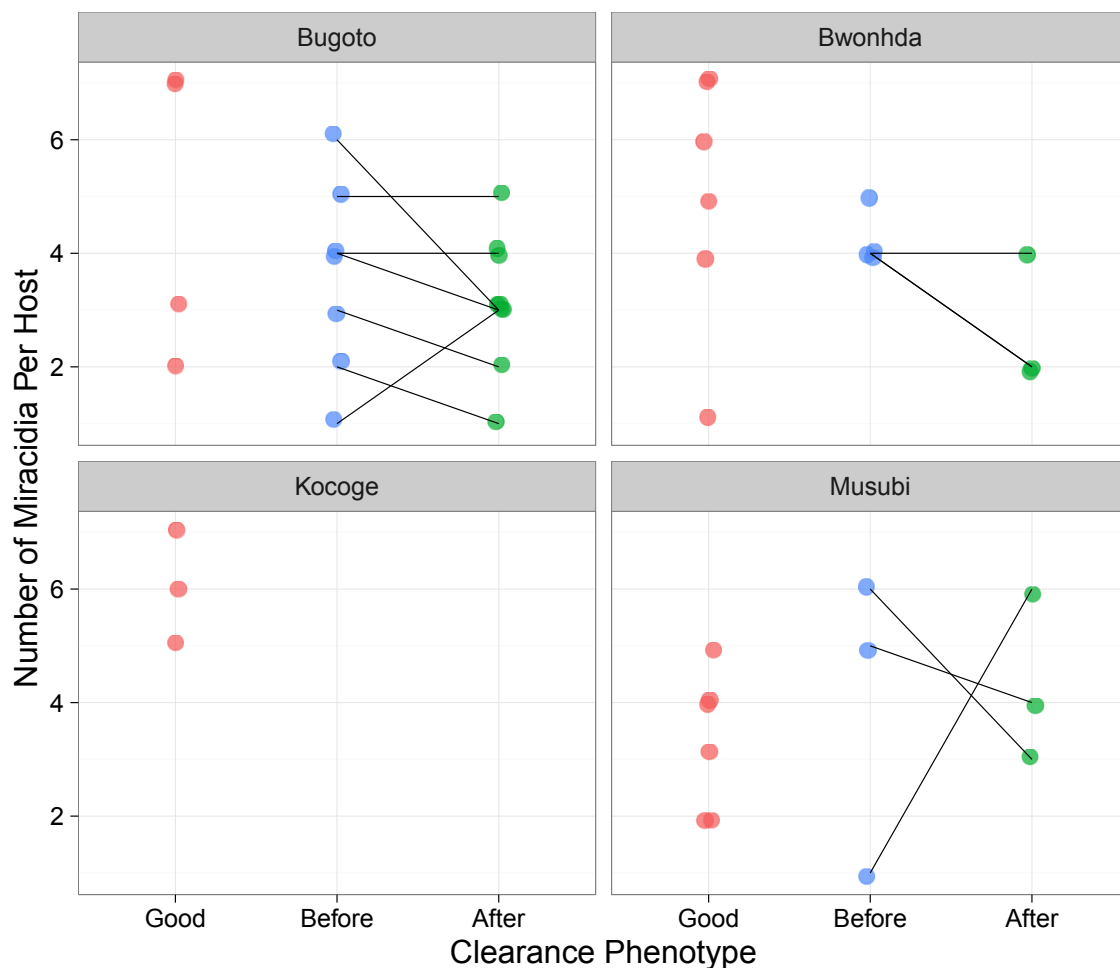


Figure 23. Number of miracidia ($n=190$) per host ($n=35$) by school and time point. Each panel represents a school and each point a sampled host at a point in time. The labels “Good” refer to miracidia from individuals that cleared completely following treatment, “Before” to pre-treatment miracidia from individuals with a poor drug response and “After” to post-treatment miracidia from individuals with a poor drug response. Miracidia that were collected from the same individual both before and after treatment are indicated by a joining line between points.

Principal component analysis of population structure

A principal component analysis (PCA) was performed on the matrix of 5.31 million variants from 190 samples using PLINK v3.40. Identity by descent values were calculated with KING v1.4 using the `--homo` algorithm which assumes a homogenous population. Kinship values were calculated between each pair of samples, values in the range of 0.177, 0.354 indicate 1st degree relatedness (i.e. parents/ children), values between 0.0884, 0.177 indicate 2nd degree relatedness (i.e. full siblings) and values between 0.0442, 0.0884 indicate 3rd degree relatedness (i.e. aunt/ uncle/ nephew/niece). Negative values indicate that individuals are unrelated [226].

Linkage disequilibrium (r^2)

To assess the extent to which the association between variants decays across the genome as a function of physical distance, values of r^2 were calculated in PLINK between pairs of SNPs within a maximum distance of 400 SNPs or 300kb. Values of r^2 lower than 0.1 were ignored. The data were plotted in R using package `ggplot2` v.2.0 and a smoothed mean and 95% confidence interval applied to the data.

Fixation index (F_{ST})

Given the nested nature of parasite sampling (we sampled multiple parasites from each individual/host, multiple hosts per school, and several schools from some districts), the data can be grouped into different hierarchical levels as we saw in the previous Chapter with the use of mixed effect models to estimate the egg reduction rate. This partitioning, or spatial structure, of natural populations has implications for the population genetic analyses. For instance, when a population deviates strongly from assumptions of panmixia, in that it is partitioned into isolated sub-populations with limited migration, this affects the probability of fixation of mutations. In general, a lower effective population size increases stochastic forces and thus increases the chance of deleterious alleles reaching fixation, though the increased homozygosity within subpopulations may result in an increased probability of fixing beneficial alleles [227]. A key statistic to assess the extent to which the variance of allelic frequencies within a population is explained by substructure is F_{ST} , which can be thought of as the proportion of genetic diversity due to allele differences among populations or as the correlations between alleles within sub-populations relative to the total population [228]. Although there are no absolute guides on how to interpret F_{ST} statistics,

Hartl and Clark (1997) have suggested the following classifications <0.05 = little genetic differentiation, $0.05-0.15$ = moderate genetic differentiation, $0.15-0.25$ = great genetic differentiation and > 0.25 very great genetic differentiation [229].

Values of F_{ST} were calculated in PLINK (v3.40) setting the subpopulations using the `--within` flag at the level of hosts, schools or districts. The variants were pruned to remove any SNPs with an r^2 value of greater than 0.3 (as this is approximately the background rate of LD for markers as far as 50 kb apart, see Figure 26) leaving a dataset of 2.11 million SNPs from 190 miracidia. The software calculates F_{ST} for each variant, a global mean across all variants and a weighted mean accounting for the sample size and number of SNPs. Any negative values of F_{ST} are spurious and so were removed from further analysis. A custom python script was used to bin values of F_{ST} every 100 SNPs and calculate the median.

Nucleotide diversity (π)

Nucleotide diversity (π ; π) was calculated on miracidia from poor clearers before ($n= 54$) and after ($n=49$) treatment using vcftools v0.1.12 in bins of 1kb. The data were split by chromosome and visualised as boxplots in R with package the ggplot2 v2.0 [230]. Confidence intervals were calculated by bootstrapping from the samples before and after treatment with 100 replicates, then taking the 5th and 95th percentile of the resulting distributions of median π values. A significant difference in π between chromosomes or time points was defined as non-overlapping 95% confidence intervals.

Genome-wide association study (GWAS)

Variants along the genome may be predictive of a binary phenotype in a case-control study or a quantitative trait. Genome-wide association studies (GWAS), in which genome-wide variants are regressed against a phenotype, have proved a powerful tool in human genetics though typically they require large sample sizes and the underlying population structure must be adjusted for to prevent confounding [231]. In this study we have sequenced miracidia from individuals that have shown different clearance rates as measured by the ERR and also from poor clearing individuals from before and after treatment. It is unclear how effectively a phenotype from the host will correlate with parasite genetic traits as each host is likely to contain multiple *S. mansoni* infections and so there is parasite genetic variation within each host [101, 232]. Further complicating the

analysis, the genotypes come from offspring (miracidia) rather than adult worms. In addition, as the effect size and alleles frequencies are not known *a priori*, I was unable to calculate the sample size required to reach genome-wide significance in the context of multiple testing [233]. Despite these limitations, correlating variants under selection with a clearance phenotype or ERR will be an important measure to determine if selection is acting due to pressure from treatment with praziquantel as opposed to other sources of selection, such as the immune systems of the intermediate or definite hosts.

To determine if any *S. mansoni* variants are associated with host parasite clearance, a logistic regression genome-wide association study was performed in PLINK, comparing cases (post-treatment miracidia, $n= 49$) with controls (pre-treatment miracidia, $n=54$). In addition to the filters imposed on all analyses, SNPs with extreme Hardy-Weinberg equilibrium values (exact test p value $< 1 \times 10^{-50}$) and with r^2 values > 0.4 within 5 kb blocks were removed, leaving 2.11 million variants, when the analysis was performed only on SNPs within chromosomes (as opposed to scaffolds in the V6 reference genome) this reduced the number of variants further to 1.85 million SNPs. Given the absence of structure between samples from before and after treatment within individuals, as seen in the PCA (Figure 24), no covariates for population structure were included.

The SNPs in the highest percentile of scores for the iHS (see below) and genome-wide F_{ST} values were also placed into a case-control regression in PLINK to examine their association with host clearance phenotype after pruning SNPs with high linkage disequilibrium ($r^2 > 0.3$ within windows of 1 kb). Here phenotype was denoted as cases (poor clearers; before and after treatment) and controls (good clearers). This phenotype was chosen above the ERR or a three factor categorical trait (good clearers, poor clearer; before treatment, poor clearers; after treatment) as i) the ERR is not normally distributed amongst our samples, which is typically required for quantitative trait GWAS [234], and contains outliers that may bias the analysis ii) the three factor categorical trait may be spurious as analysis shows there to be little differentiation between miracidia from before and after treatment (see below). The significance threshold to correct for multiple testing is a false discovery rate (FDR) of < 0.05 , this correction reduces the number of type I errors, or false positives, to a rate equal or less than the FDR. For instance, given a cut off of 0.05, fewer or equal to 5% of the genes with this FDR are expected to be false positives.

Integrated haplotype score (iHS) and F_{ST} to detect positive selection

When analysing F_{ST} along whole genomes the values of neutral loci will be determined by the demographic history of the population, therefore outlier regions showing unusually large amounts of differentiation may be indicative of positive selection [141, 228]. The integrated haplotype score (iHS) calculates the decay in the relatedness of SNPs along a haplotype as a function of distance for both ancestral and derived alleles and then calculates the integral of the observed decay in relatedness in both directions before dividing the score of the ancestral allele over the derived allele [235].

To calculate the iHS it was necessary to provide a recombination map. While no detailed map exists for *S. mansoni*, I created a genetic map under the (simplistic) assumption that recombination was constant across the genome at a rate equal to that reported from microsatellite data of 227.2 kb/cM (95% CI 181 to 309 kb/cM) [236]. The variants were split into 7 autosomal chromosomes and statistically phased using SHAPEIT v2.r837 [237]; the model was run for 50 MCMC iterations within windows of 0.5 mb as recommended in the manual. The effective population size was set at 65,000 based on the previous estimate for the Ugandan population in Chapter 2. Phased VCFs were then used in the software Selscan v1.1.0 [238] to calculate iHS values for the SNPs in each chromosome separately before combining the chromosomes. A custom python script was used to bin every 100 SNPs and calculate the median within each bin. As the iHS is not designed to detect selection on rare variants a more stringent MAF filter of 0.05 was applied, reducing the total number of SNPs analysed to 1.78 million variants.

Locating causative SNPs

Single nucleotide and indel variations overlapping with the coordinates of the calcium/ sodium channel exons in the window of selection identified through iHS and F_{ST} scans of the genome were discovered using bedtools v2.17. The indels were called with GATK HaplotypeCaller, though for maximum sensitivity no QC filtering was applied at this stage. The variants were then annotated using SnpEff v4.3b [239] to denote whether they were synonymous or non-synonymous (missense) mutations and to estimate their effect on protein function. Finally a MAF filter of 0.1 was applied in PLINK to remove rare variants. These variants were subsequently associated with the binary host clearance phenotype to test for association with reduced *S. mansoni* clearance following treatment with praziquantel, using a linear model in R.

Expression data from S. mansoni genes of interest

The RNA-seq expression data for calcium channel genes in the window of selection comes from a laboratory strain of *S. mansoni* NMRI Puerto Rico; an *in vivo* infection model was set up at the University of Nottingham and the worms were removed by portal perfusion at different time points following infection (days 13, 17, 21, 28, 35) or by lung incubation at for day 6 worms. RNA was extracted using TRIzol coupled with RNA Clean & Concentrator-5 (Zymo Research) and Truseq stranded RNA libraries prepared for sequencing on the Illumina HiSeq with 75bp paired end reads. Reads were then mapped to the version 5.2 *S. mansoni* reference genome and expression data analysed using DESeq2. 'Count' refers to estimated size factor normalised counts calculated from DESeq2.

Results

Population structure

The PCA shows that the largest contributing factor to the structure of the SNP data is variance between districts (PC1 explains 10.8% of variance), as schools Musubi, Buguto and Bwonhda are located in Mayuge District whereas Kocoge is located ~100km east in the district of Tororo. Within Mayuge, miracidia from the three schools were undifferentiated, indicating that in schools around Lake Victoria the parasite population is unstructured as a consequence of high gene flow between the communities from which parasites were sampled. The second principal component (explaining 6.3% of variance) is correlated with the sequencing depth achieved for the samples. A linear regression between PC2 and sample depth of the miracidia showed a negative correlation (coefficient = -0.0086, $p=2 \times 10^{-16}$) and the R-squared value = 0.50, indicating that half of the variance in PC2 can be explained by the sample depth. This therefore represents 'technical variation' as a result of the variable coverage between samples (Figure 24), which has presumably impacted on rates of SNP calling. Varying the filtering parameters in GATK and PLINK failed to remove this aspect of variance. Individual miracidia from within individuals also did not cluster by individual host along any PCs, suggesting that there is as much diversity within individuals as in the overall population. The PCA also reveals that there is no separation by clearance type (good, poor: before, poor: after) along any of the initial PCs (Figure 24, panel B). This suggests that host

parasite clearance host does not have a major effect on the structure of the miracidia population as detected by genome-wide SNPs.

With the inclusion of sequence data from adult worms from Cameroon, Lake Victoria and Lake Albert (introduced in Chapter 2) the first principal component (13.9% of variance) separated Cameroon from the rest of the samples, which all originated in Uganda (Figure 25). The samples within Uganda were separated along the second and third principal components (explaining 7.5% and 4.9% of variance respectively), which distinguishes the population from Lake Victoria (miracidia and adult worm), the miracidia from Kocoge and the adult worm from Lake Albert in the west of Uganda.

Pairwise values of identity by descent (IBD) ranged between -0.29470, 0.11380 (median - 0.09820). The highest five pairwise values of IBD suggest 2nd order relationships (there are no 1st order relationships), 3 of these pairs are within the same host and time point, which indicate half sibling relationships. To determine if there is an effect of host, school or district on the IBD scores between miracidia a linear model was constructed and fitted in R. The IBD between a pair of samples was regressed against binary covariates noting if the pair of miracidia came from the same host, school or district. The host covariate also interacts with time, (i.e., whether the samples come from the same time point; before or after treatment). The linear model is defined as:

$$\text{IBD} \sim \text{host*time} + \text{school} + \text{district} \quad (4)$$

The covariate with the largest effect on IBD is district, i.e., samples within the same district had significantly higher IBD than between districts (coefficient = 0.023, $p = 2 \times 10^{-16}$). The effect of district is confounded with time (as all the samples from Tororo district are from 'good' clearers before treatment), therefore the analysis was subsequently restricted to pairs of samples from before and after treatment in poor clearing individuals (hence only from schools in Mayuge district). The modified linear model is defined as:

$$\text{IBD} \sim \text{host*time} + \text{school} \quad (5)$$

In this model, pairs of parasites sharing a host had a slightly higher kinship coefficient score (coefficient= 0.0097, $p= 0.018$). Miracidia from the same time point (before or after treatment with praziquantel) are also more related (coefficient = 0.0018), though the effect is small and not significant ($p=0.17$). The interaction term between host and time (i.e. parasites come from the same host and time point) had a positive coefficient (0.0041), though again the effect was not significant ($p= 0.47$). Surprisingly, miracidia from the same school were found to be *less* related than miracidia from different schools (coefficient = -0.010, $p= 3.2 \times 10^{-12}$). The IBD scores between pairs of parasites therefore confirm the finding from the PCA that there is no clear structuring of parasites within schools, though there is an effect of district.

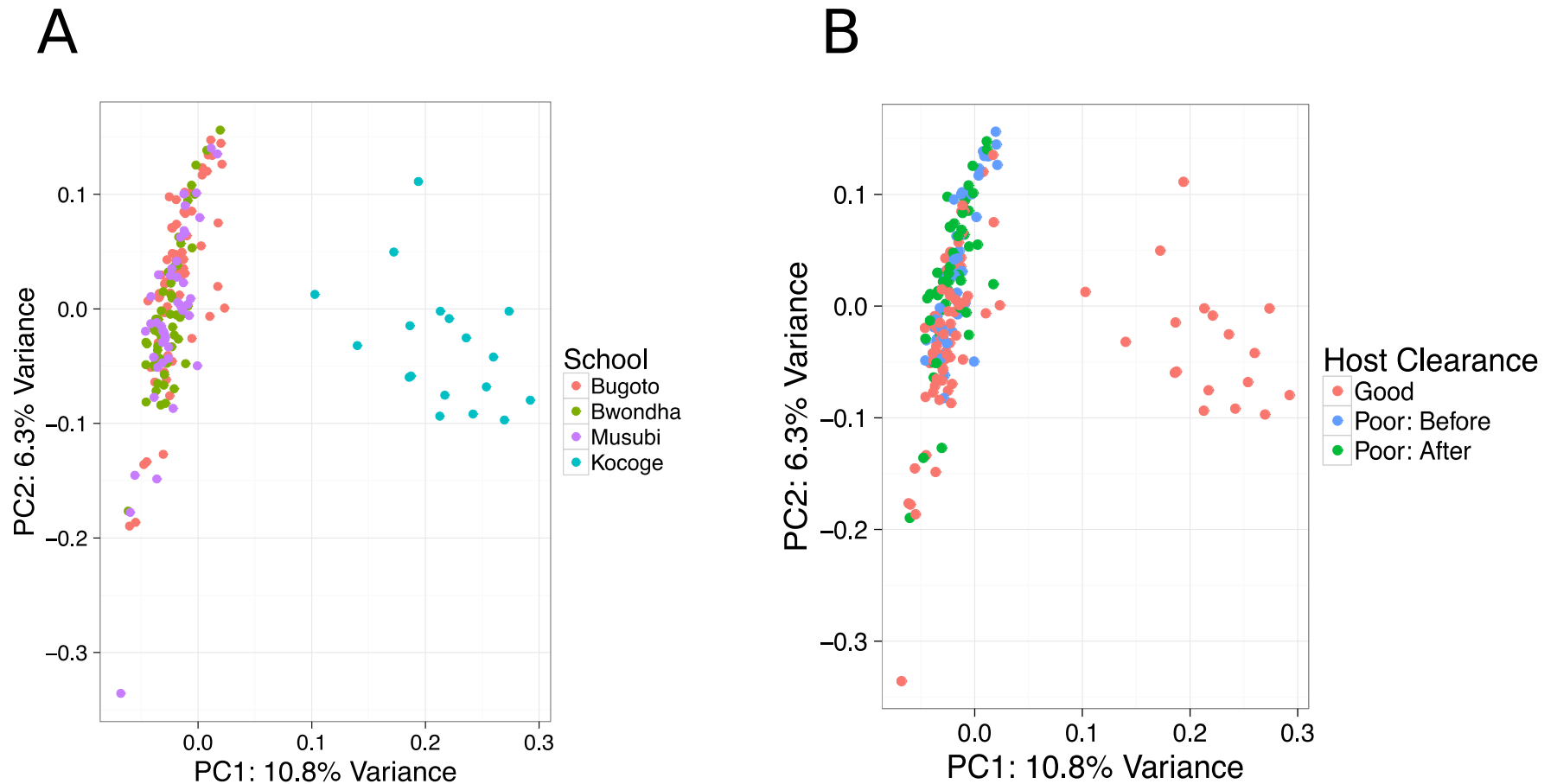
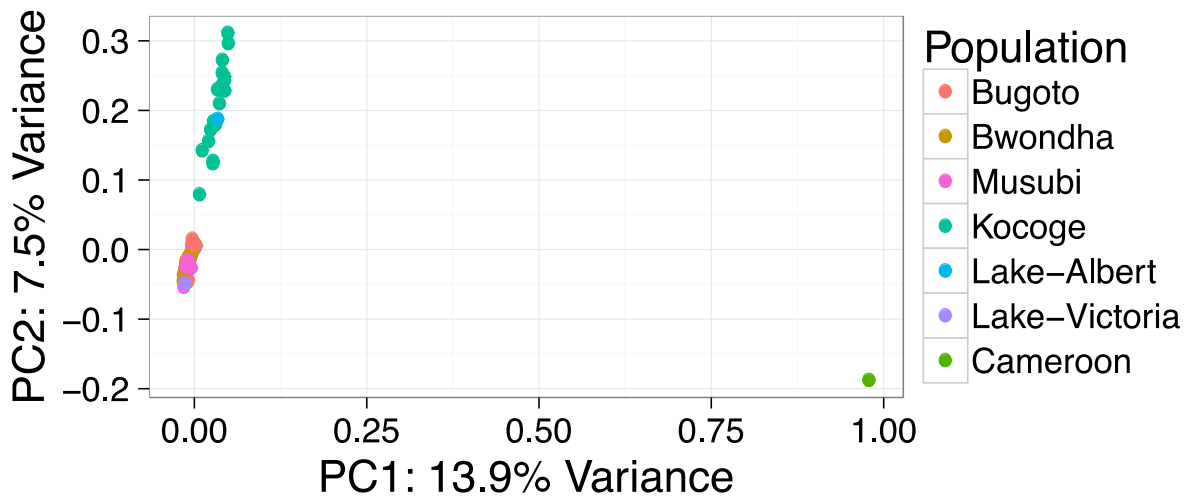


Figure 24. Principal components analysis of 5.31 million filtered SNPs from 190 *S. mansoni* miracidia collected in Uganda.

The first PC shows population structure between the schools from Mayuge District (Bugoto, Bwondha, Musubi) and the school from Tororo District (Kocoge). The second principal component is correlated with the depth of the miracidia and so represents a technical effect. The miracidia are coloured by school in panel A. The miracidia are coloured by host clearance phenotype in panel B, note that there is no clear structuring of miracidia along the first two principal components by host phenotype.

A



B

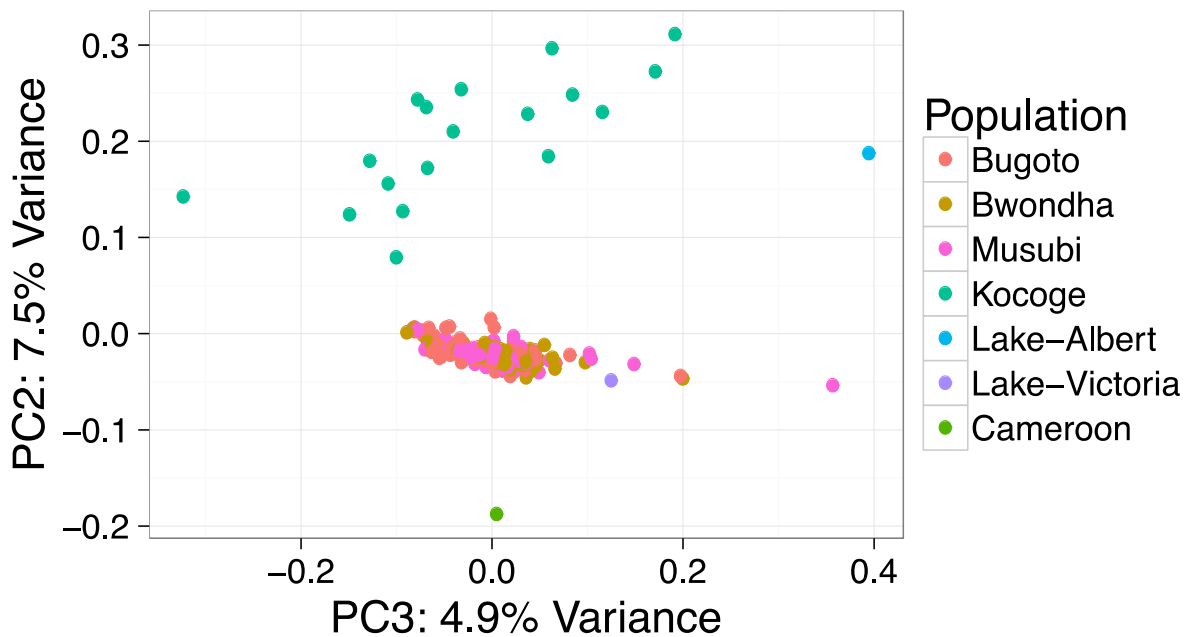


Figure 25. Principal components analysis using 15.5 million SNPs from 190 Ugandan miracidia along with isolates Lake-Albert, Lake-Victoria and Cameroon sequenced from adult worms in Chapter 2.

Panel A shows principal components (PC) 1 and 2 and Panel B shows PC 2 and 3.

Decay in linkage disequilibrium

The values of r^2 are shown between pairs of genotypes up to a distance between variants of 50 kb in Figure 26. While r^2 was calculated up to a possible maximum 300 kb (see Methods) beyond 50 kb there are fewer paired comparisons in this analysis and thus fewer data points thereby increasing the uncertainty in the data. The smoothed average shows that r^2 declines rapidly over the first 1 kb to ~ 0.4 and thereafter falls more gradually until it stabilises at ~ 0.3 beyond physical distances of >10 kb. Subsequent fluctuations at distances >10 kb are most likely attributed to noise in the data.

Fixation index

The values of F_{ST} were calculated between hosts ($n=35$), schools ($n=4$) and districts ($n=2$) and are output as a global mean or a weighted estimate based on the sample size [240] and are shown in Table 9. Overall these values showed there was very little structure within the parasite population at any spatial resolution; all values fall below 0.05, defined by Harl and Clark (1997) as representing “little” genetic differentiation [229]. The greatest partitioning of variation in this dataset occurs between districts (mean $F_{ST} = 0.0232$), mirroring the trend seen in the PCA and with IBD relationships in which the district the parasite was sampled from has an effect on population structure but the school does not.

Comparison of miracidia population diversity before and after treatment within hosts

To determine if a single round of treatment had an effect on the variability of the parasites before and after treatment, nucleotide diversity (π) was calculated for 54 samples before-treatment and 49 samples after treatment from hosts showing a sub-optimal drug response. There is variance between chromosomes, with the ZW sex chromosome showing the least diversity (median 1.47×10^{-3}), this difference is to be expected as the sex chromosome largely represents Z (W specific regions are fragmented and not included in the analysis); therefore female miracidia in our sample will show reduced diversity in this region as reads from the W chromosome will not map to this chromosome. Among the seven autosomal chromosomes the median π is similar, though chromosome 5 shows significantly more diversity than other autosomal chromosomes (median 3.09×10^{-3}).

When comparing between time points, the genome-wide values of π from before (median 1.87×10^{-3} ; 95% CI 1.77×10^{-3} , 1.98×10^{-3}) and after (median 1.97×10^{-3} ; 95% CI 1.88×10^{-3} , 2.06×10^{-3}) treatment show overlapping 95% confidence intervals as calculated by bootstrapping and these values show little variation when subset by chromosome (Figure 27). Therefore there is no evidence in my data to suggest that there is a reduction in *S. mansoni* genetic diversity following a single round of treatment with praziquantel in an area exposed to ≥ 8 previous rounds of MDA.

Calculating F_{ST} between miracidia from before ($n = 54$) and after ($n = 49$) treatment, the level of differentiation between these two groups is very low (mean $F_{ST} = 0.0000251$, weighted $F_{ST} = 0.000185$), suggesting little difference in allele frequencies between them. Figure 28 shows how the F_{ST} between before and after treatment samples varies along the genome; there are certain areas that show F_{ST} values well above the background level though no windows of 100 contiguous SNPs have a median F_{ST} above 0.1.

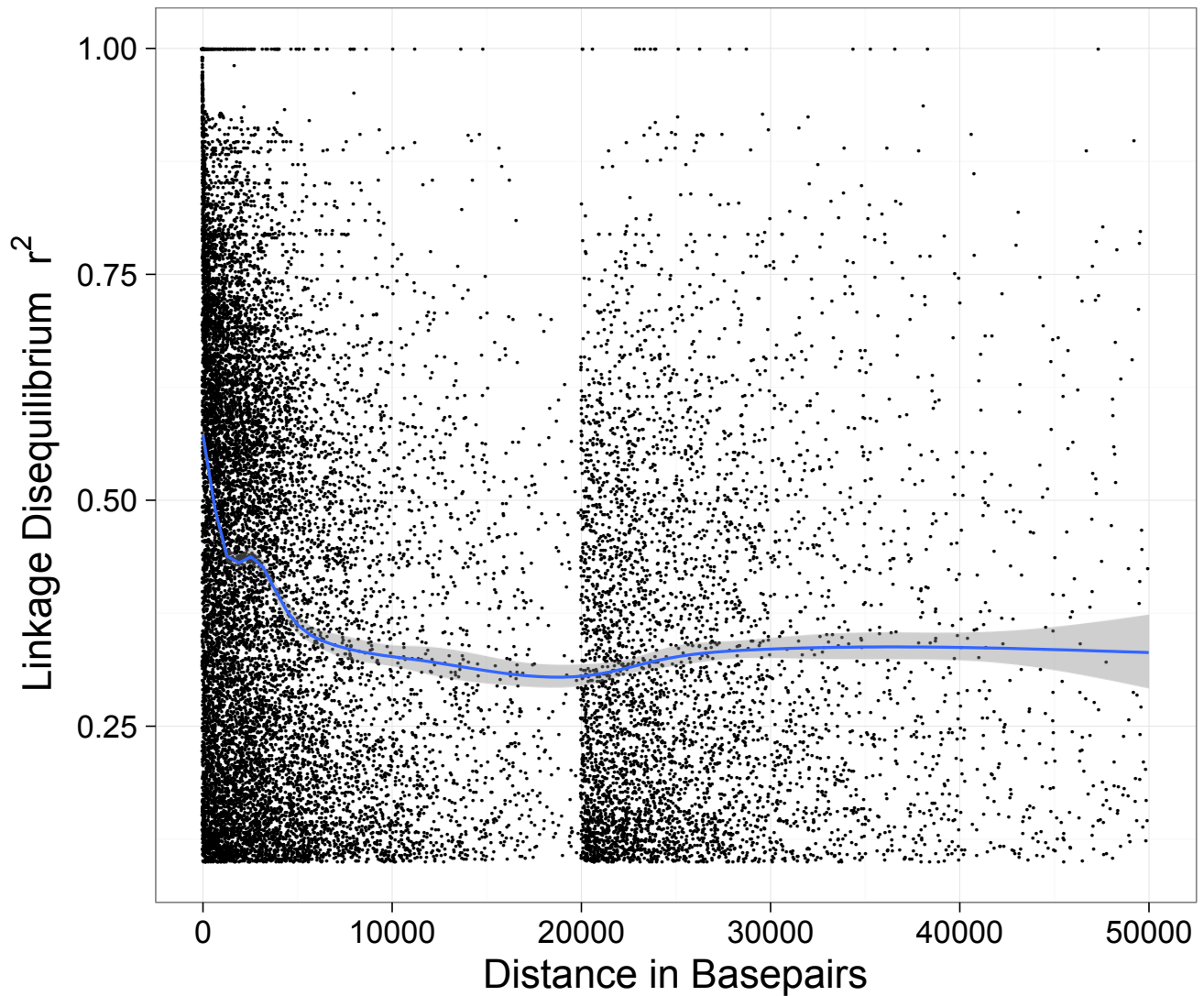


Figure 26. Linkage disequilibrium as measured by r^2 between pairs of genotypes within chromosomes.

A smoothed average (blue line) has been plotted through the points with a 95% confidence interval (grey shading). The average r^2 declines sharply within the first 1 kb of physical distance to and continues to decline until around 10 kb, where it remains largely constant at ~ 0.3 for markers as far as 50 kb apart. Calculating r^2 between chromosomes (not shown) resulted in a median r^2 of ~ 0.2 , suggesting that r^2 may decline further if plotted over longer physical distances than 50 kb.

Table 9. Mean F_{ST} values at different spatial scales.

Values are calculated using 2.11 millions SNPs from 190 *S. mansoni* miracidia collected in Uganda at the level of individual individuals, schools and districts. The variants are in approximate linkage equilibrium as SNPs with $r^2 > 0.3$ in windows of 5 kb have been pruned.

F_{ST}	Between individuals (35 clusters)	Between schools (4 clusters)	Between districts (2 clusters)
Mean	0.0109	0.00800	0.0232
Weighted	0.0105	0.00812	0.0301

Table 10. Nucleotide diversity values by chromosome and time point from miracidia collected in Uganda before ($n=54$) and after ($n=49$) treatment with praziquantel.

Median values and 95% confidence intervals are presented.

Chromosome	Before treatment	After treatment
All	1.87x10 ⁻³ (1.77x10 ⁻³ , 1.98x10 ⁻³)	1.97x10 ⁻³ (1.88x10 ⁻³ , 2.59x10 ⁻³)
1	1.64x10 ⁻³ (1.54x10 ⁻³ , 1.74x10 ⁻³)	1.71x10 ⁻³ (1.63x10 ⁻³ , 1.79x10 ⁻³)
2	2.34x10 ⁻³ (2.20x10 ⁻³ , 2.47x10 ⁻³)	2.52x10 ⁻³ (2.39x10 ⁻³ , 2.63x10 ⁻³)
3	1.82x10 ⁻³ (1.71x10 ⁻³ , 1.93x10 ⁻³)	1.89x10 ⁻³ (1.80x10 ⁻³ , 1.98x10 ⁻³)
4	1.93x10 ⁻³ (1.81x10 ⁻³ , 2.05x10 ⁻³)	2.01x10 ⁻³ (1.92x10 ⁻³ , 2.12x10 ⁻³)
5	3.02x10 ⁻³ (2.86x10 ⁻³ , 3.18x10 ⁻³)	3.15x10 ⁻³ (2.99x10 ⁻³ , 3.28x10 ⁻³)
6	2.16x10 ⁻³ (2.04x10 ⁻³ , 2.29x10 ⁻³)	2.24x10 ⁻³ (2.13x10 ⁻³ , 2.35x10 ⁻³)
7	2.44x10 ⁻³ (2.31x10 ⁻³ , 2.58x10 ⁻³)	2.48x10 ⁻³ (2.38x10 ⁻³ , 2.62x10 ⁻³)
ZW	1.39x10 ⁻³ (1.31x10 ⁻³ , 1.50x10 ⁻³)	1.53x10 ⁻³ (1.46x10 ⁻³ , 1.60x10 ⁻³)

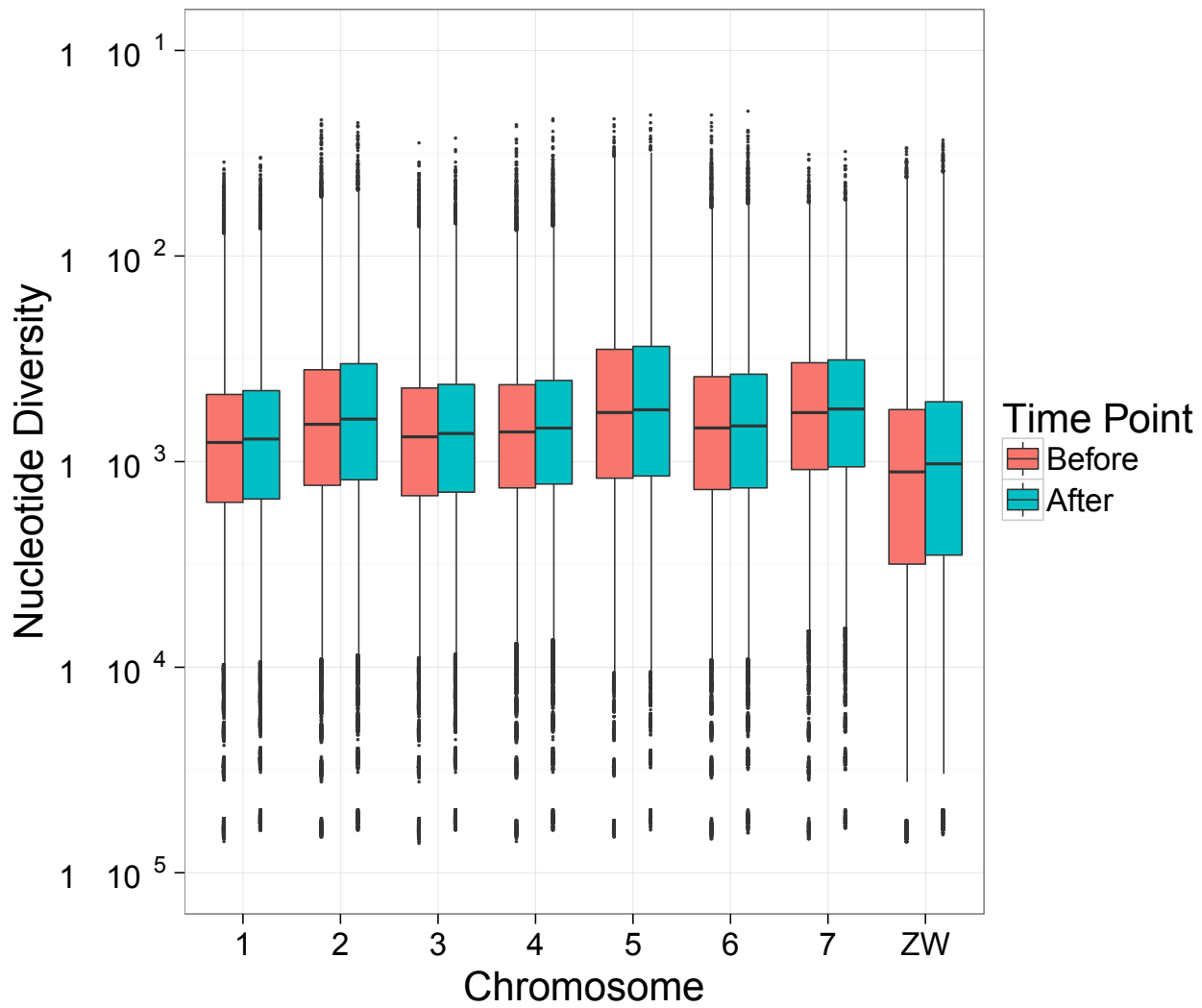


Figure 27. Values of nucleotide diversity of *S. mansoni* miracidia (π ; π) for each chromosome. Boxplots are partitioned by time point; before ($n= 54$) and after ($n=49$) treatment with praziquantel. Note that due to the long tail of π values the y-axis is presented on a log₁₀ scale.

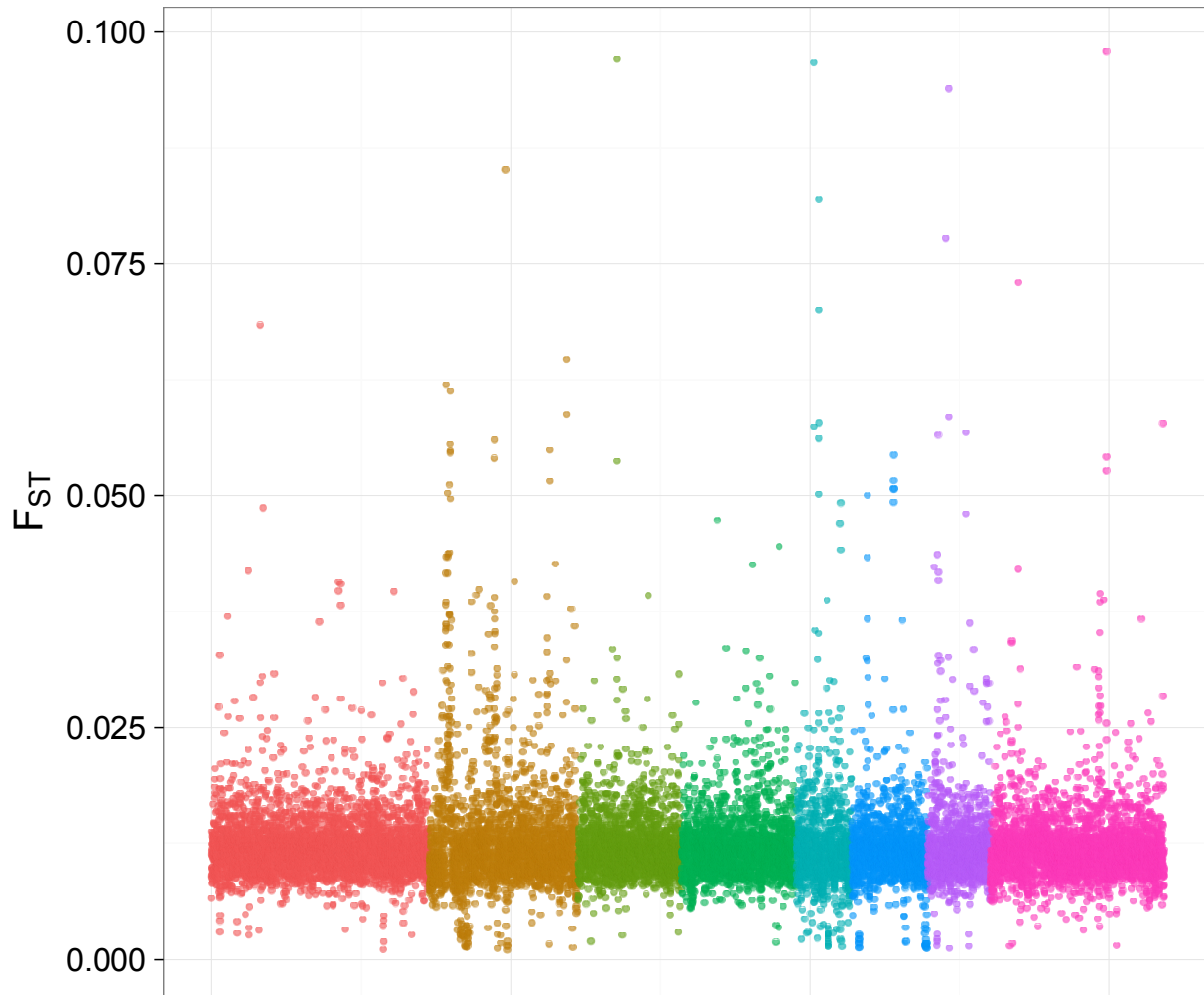


Figure 28. Values of F_{ST} comparing miracidia from before ($n=54$) and after ($n=49$) treatment with praziquantel from hosts in three intensively treated schools in Uganda.

Every point represents the median of a bin of 100 SNPs and the points are coloured by chromosome. While the mean background level of F_{ST} is very low (~ 0.0001) certain regions show higher levels of differentiation, though F_{ST} never exceeds 0.1.

Genome-wide association between clearance phenotype and SNP variants

A genome-wide association study of miracidia from before (controls; $n=54$) and after (cases; $n=49$) treatment with praziquantel from hosts with suboptimal clearance found no SNPs to be significantly associated with time point after a multiple test FDR correction. The lowest uncorrected p value was 1.68×10^{-5} , which is greater than the significance threshold for 1.85 million regressions ($p = 3.37 \times 10^{-8}$ given a Benjamini-Hochburg procedure for correcting p values to a FDR of 0.05 [241]). This indicates that either the test was underpowered to detect differences between the populations or that there are insufficient genetic differences between miracidia from before and after treatment in heavily treated populations to be detectable given our sampling method and phenotype.

Genomewide F_{ST} and integrated haplotype score (iHS)

Setting the subpopulations at the level of hosts ($n=35$), we have shown above that the overall level of differentiation is very low (mean = 0.0109, Table 9). However, when scanning F_{ST} along the genome, we can observe considerable heterogeneity in F_{ST} values along each chromosome (Figure 29). This is of interest as outlier values of F_{ST} have been shown to indicate regions under selection within the genome. It is clear that several regions are outliers well above the background level of differentiation. The two bins with the highest median values ($F_{ST} = 0.41-0.42$) are found at the beginning of chromosome 2.

The iHS has a mean value close to zero when averaged along the genome (-0.147), however when scanning along the genome considerable heterogeneity can be observed in certain regions (Figure 30), the values have been binned as the median of every 100 SNPs as iHS is more powerful to detect selection when looking in windows of consecutive SNPs carrying extreme iHS values [235]. Extreme positive values of iHS indicate long haplotypes carrying derived (or alternate) alleles and extreme negative values indicate long haplotypes carrying ancestral (or reference) alleles. The two bins with the most extreme values (iHS = 3.3) are found at the beginning of chromosome 2.

The variants with the highest values from both tests of selection were then placed into a case-control association test against the host clearance phenotype. Examining the highest percentile of binned values from F_{ST} (Figure 29; $n= 24516$ SNPs from 246 bins) and iHS (Figure 30, $n= 17800$ SNPs from 178 bins), 41617 were unique to either list (98.3%). Note that there are more values

for F_{ST} as a more stringent MAF filter (0.05) was applied to variants in the iHS test as haplotype based tests are not designed to detect selection acting on rare variation. After pruning of SNPs in close linkage, the number of variants was reduced to 6698. An association study was performed in PLINK, examining the association between the 6698 variants in the windows of highest selection against a binary clearance phenotype; miracidia were classified according to their hosts showing good or poor parasite clearance following treatment with praziquantel. This phenotype therefore does not distinguish between miracidia from before or after treatment amongst hosts with poor clearance as we have shown above that there are no significant differences between these time points. Four variants showed significant p -values after the correction for multiple testing ($FDR < 0.05$), of these 3 SNPs were found in the 256 kb region of chromosome 2 (base-pair coordinates: 671654, 678447, 678618; uncorrected p -values: 1.83×10^{-5} , 1.86×10^{-5} , 6.33×10^{-6}) and a fourth on chromosome 4 (basepair coordinates: 3007159, uncorrected p -value: 2.76×10^{-5}), see Figure 31.

Both the genome-wide F_{ST} plot and iHS indicate that a region at the beginning of chromosome 2 is under positive selection. The variants with the highest values from both tests span a 256 kb region at the beginning of chromosome 2 (basepair coordinates: 659,623-915,593). Although the high values in the F_{ST} and iHS tests suggest that a hard sweep may be occurring, we previously found no reduction in nucleotide diversity in chromosome 2 from before and after treatment (Table 10, Figure 27), which would be indicative of a selective sweep. Therefore to determine if the alternative allele in the window of selection is associated with miracidia from a particular district or time point, a multinomial logistic regression was performed in R v3.2.2 with package nnet v7.3-10. Initially the haplotypes of the 12774 phased variants in the 256 kb region from 190 miracidia were associated against district. The odds ratios of a heterozygous allele on either haplotype (0|1, or 1|0) or homozygous alternative allele (1|1) in miracidia from Mayuge district (Bugoto, Bwondha or Musubi schools) were 3.43 (95%CI 3.35, 3.51), 3.51 (3.43, 3.59) and 8.90 (8.73, 9.08) respectively. By contrast, when restricting the analysis to miracidia collected from hosts before ($n=54$) or after ($n=49$) treatment with praziquantel (therefore only schools in Mayuge), the odds of the heterozygous haplotypes or the homozygous alternative haplotype in miracidia after treatment were 0.922 (95%CI 0.892, 0.953), 0.947 (0.916, 0.979) and 1.01 (0.979, 1.04) respectively. These differences in odds ratios show that the effect of district is much greater in determining the haplotype of the miracidia in this 256 kb region (odds ratios ≥ 3.4) than is the time point (odds ratios ≈ 1). This suggests that the variable haplotype lengths are driven by

differences between the districts (schools in Mayuge had 8 or 9 previous rounds of MDA with praziquantel compared to 1 previous round in Tororo, see Chapter 3), rather than differences in parasites surviving a single round of treatment within Mayuge.

Although the annotation for this version of the reference genome is not complete, it is predicted that within the 256 kb region on chromosome 2 are 4 genes, of which 2 contain pfam domains for sodium/ calcium channels (Table 11).

The SNP in chromosome 4 that was significantly associated with the clearance phenotype (Figure 31) is within 20 kb upstream and downstream of three predicted genes, two of these (g6782, g6784) contain protein domains annotated as Leishmanolysin, a membrane-bound glycoprotein found on the promastigote of species of the protozoan parasite *Leishmania* [242]. The other predicted gene close the SNP of interest (g6783) contains no known protein domains.

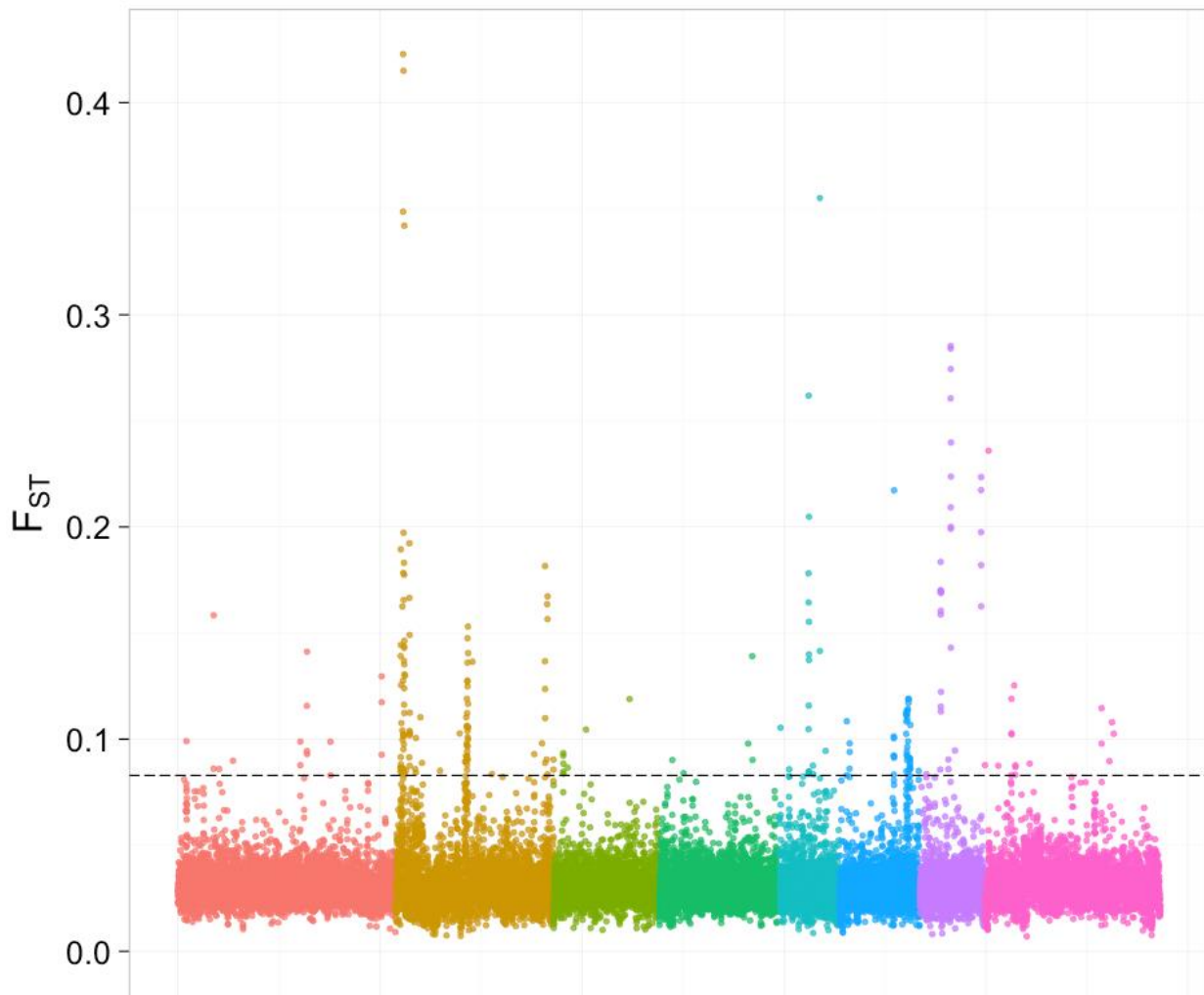


Figure 29. Values of F_{ST} plotted along the *S. mansoni* genome from 190 miracidia. The subpopulations are within hosts ($n=35$).

While the background level of F_{ST} is low (mean = 0.0115), a number of outlier regions are apparent which may be indicative of positive selection. Each point represents the median F_{ST} from a bin of 100 SNPs, total bins = 24315. The dotted line indicates the top percentile of SNPs ($F_{ST} > 0.0829$), above which are 246 bins containing 24516 SNPs. The SNPs within these bins were used for the association study against host clearance phenotype shown in Figure 31. Note that some bins may contain fewer than 100 SNPs if they are found at the end of a chromosome.

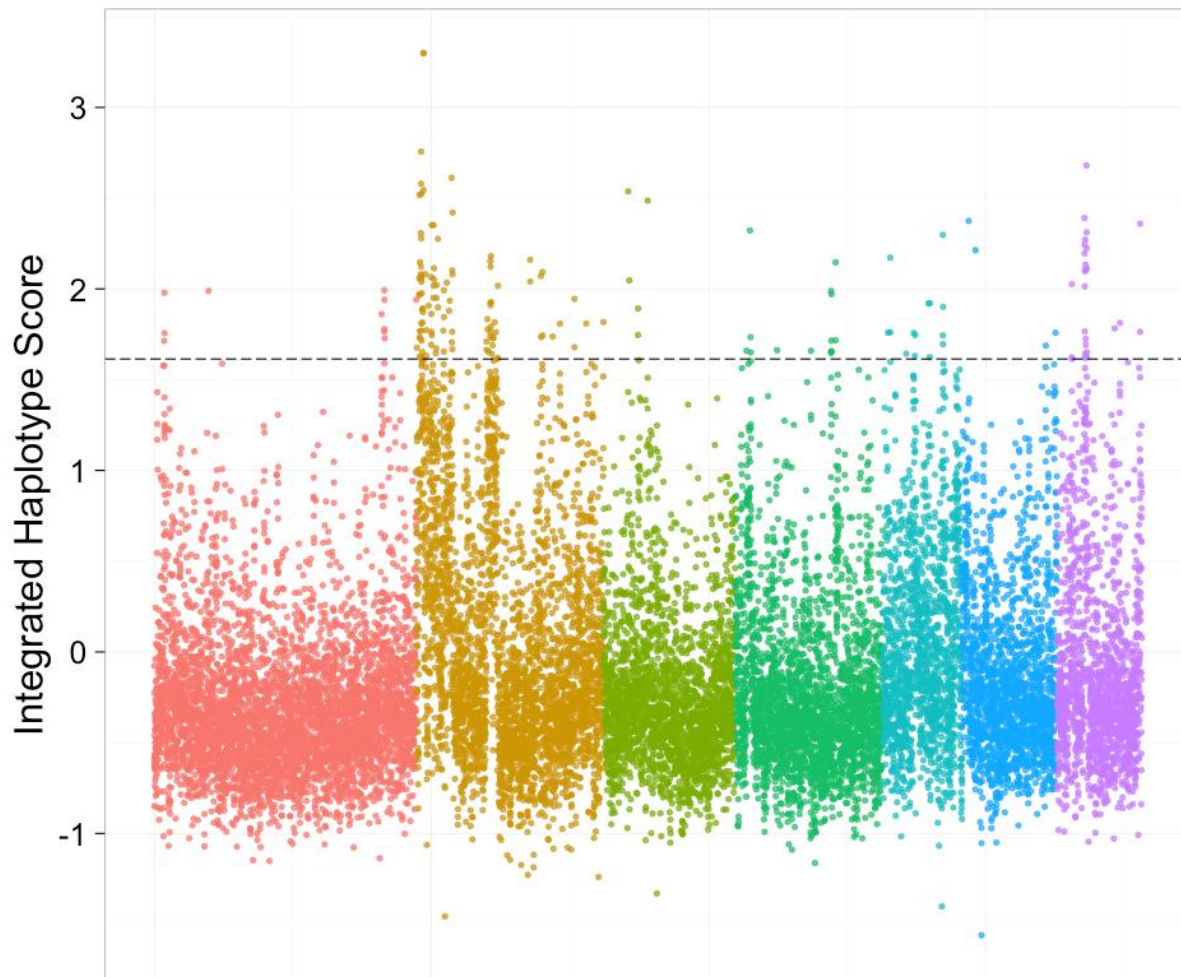


Figure 30. The integrated haplotype score (iHS) plotted along the 7 autosomal chromosomes of the *S. mansoni* genome from 190 miracidia.

Extreme positive values of iHS indicate long haplotypes carrying derived alleles. While the background level of iHS is low (mean = -0.147), several outlier regions can be detected. Each point represents the median iHS from a bin of 100 phased SNPs, total bins = 17805. The dotted line indicates the top percentile of values (iHS > 1.614), above which are 178 bins containing 17800 SNPs. The SNPs within these bins were used for the association study against host clearance phenotype shown in Figure 31.

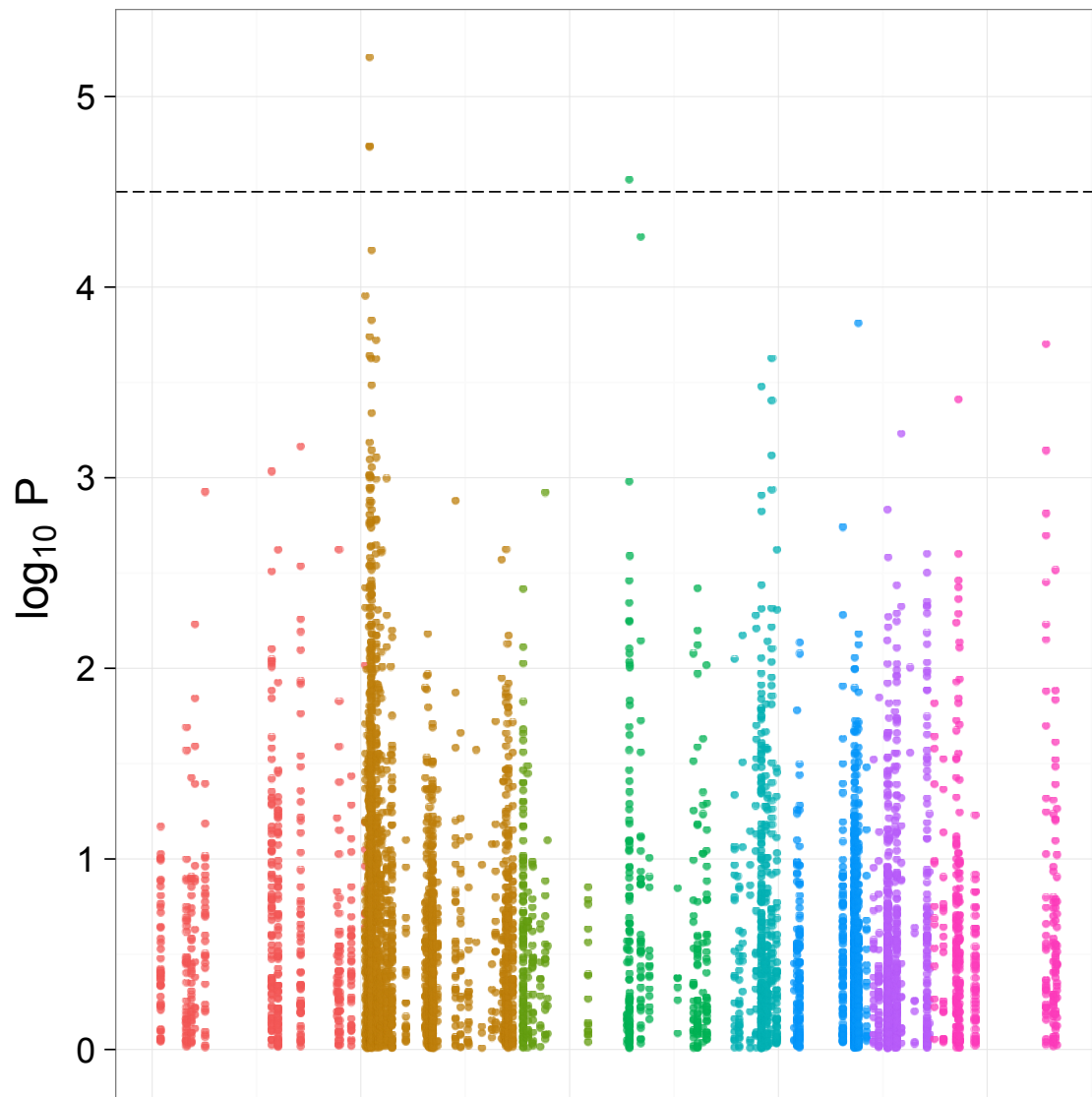


Figure 31. Case control association study on the top 1% of SNPs under selection in 190 *S. mansoni* miracidia after removing markers in close linkage (6698 variants).

The test compares miracidia from ‘good clearers’ (control; $n= 87$) against miracidia from ‘poor clearers’ (cases; $n= 103$) from 35 hosts from four primary schools in Uganda. Each point represents a single variant and points are coloured by chromosome. The four variants above the hashed line have $FDR < 0.05$ (the threshold for significance after correction for multiple testing); three are within the 256 kb window of selection at the beginning of chromosome 2 and one is located in chromosome 4.

Table 11. Genes from a 256 kb region of selection on chromosome 2 of the V6 *S. mansoni* reference genome. Coordinates indicate the location of the gene in kilobases.

Pfam domains refer to matches within the protein family database of the EBI. Note that g4218 has no matching ID from the V5 *S. mansoni* reference genome, nor does it contain any recognised pfam domains. See Table 12 for information on the non-synonymous SNPs within the two genes with predicted calcium channels.

Gene ID (V6/V5)	Coordinates (kb)	Number of exons	Pfam domains	Function
g4216 Smp_094380	658-679	10	PHD finger, FHA domain	Zinc finger proteins involved in chromatin-regulated gene expression; Forkhead associated domains are present in many regulatory proteins
g4217 Smp_170370	794-799	4	Sodium/ Calcium exchanger protein	Regulates intracellular Ca ²⁺ concentration depending on Na ⁺ concentration
g4218	846-847	1	NA	NA
g4219 Smp_170450	901-947	9	Sodium/ Calcium exchanger protein (x2)	Regulates intracellular Ca ²⁺ concentration depending on Na ⁺ concentration

Locating causative SNPs

Within the exons of the calcium channel genes located in the 256 kb window of selection there are 75 SNPs. Of these, 41 are annotated as missense variants and 4 have a minor allele frequency > 0.1 (Table 12). There are also two indels inside the exons with a MAF > 0.1 across all samples (Table 12). In examining these 6 non-synonymous polymorphisms more closely, all SNP variants are predicted to have at least 'moderate' impact on the function of the protein, indicating that while these non-synonymous changes are not disruptive, they may change protein effectiveness. Both indels cause frameshift mutations, in which entire amino acids are skipped resulting in protein truncation, which are predicted to result in 'high' impact to protein function. Interestingly, the two indels are adjacent to one another in the genome and occur at a high MAF (~0.4), suggesting that the frameshift mutations are not deleterious to *S. mansoni* in western Uganda or the variants would not be able to exist at such a relatively high frequency. Associating the alleles of the 6 variants against the binary host clearance phenotype in a generalised linear model, only the first variant (position 793721 in predicted gene g4217; Table 12) was found to show a significant association with host clearance given a threshold of $p < 0.05$ (heterozygote coefficient = -3.613, $p = 0.00743$; homozygous alternative coefficient = -4.809, $p = 0.00260$).

Expression of calcium channel genes

The strongest signal of positive selection and association are located in a region that includes calcium channel genes. Therefore data on the expression of these genes would be valuable to confirm that the genes are expressed during the adult stage of the *S. mansoni* lifecycle as this is the stage against which praziquantel is effective. We would also expect the causative genes to show less or no expression during the juvenile schistosomula stage as praziquantel is less effective against immature parasites. A separate experimental study has provided evidence on the time course of expression of these two genes within the 256 kb region (Smp_170370 and Smp_170450; see Table 11). As the parasites are maturing into adult worms the normalized counts of these genes rose substantially 6 and 13 days following infection in an *in vivo* mouse study (see Figure 32). This therefore shows that the genes are expressed during the adult life stage and that they show a higher level of expression in adults than juvenile worms, which is consistent with the differences in drug response observed between the life stages.

Table 12. Non-synonymous variants and indels within the exons of calcium channel genes inside the 256 kb window of selection on chromosome 2.

The variable amino acids and predicated damage to protein structure have been estimated through the snpEff program given the annotation available for these genes. Variants with 'high' predicted impact are assumed to cause truncation or loss of function. Variants with 'moderate' impact are predicted to be non-disruptive though they may change protein effectiveness. Note FS indicates a frameshift mutation. The minor allele frequency (MAF) is also provided for each gene.

Variant Position	Type	Alleles	Gene	Amino Acids	MAF	Variant impact
793721	SNP	A,G	g4217	Lys, Glu	0.110	Moderate
908752	SNP	G,T	g4219	Leu, Phe	0.184	Moderate
908855	SNP	A,T	g4219	Asn, Tyr	0.384	Moderate
912307	SNP	A,C	g4219	Glu, Asp	0.258	Moderate
912264	Indel	ATCTGT, A	g4219	Leu, FS	0.390	High
912271	indel	ATC, A	g4219	Ser, FS	0.403	High

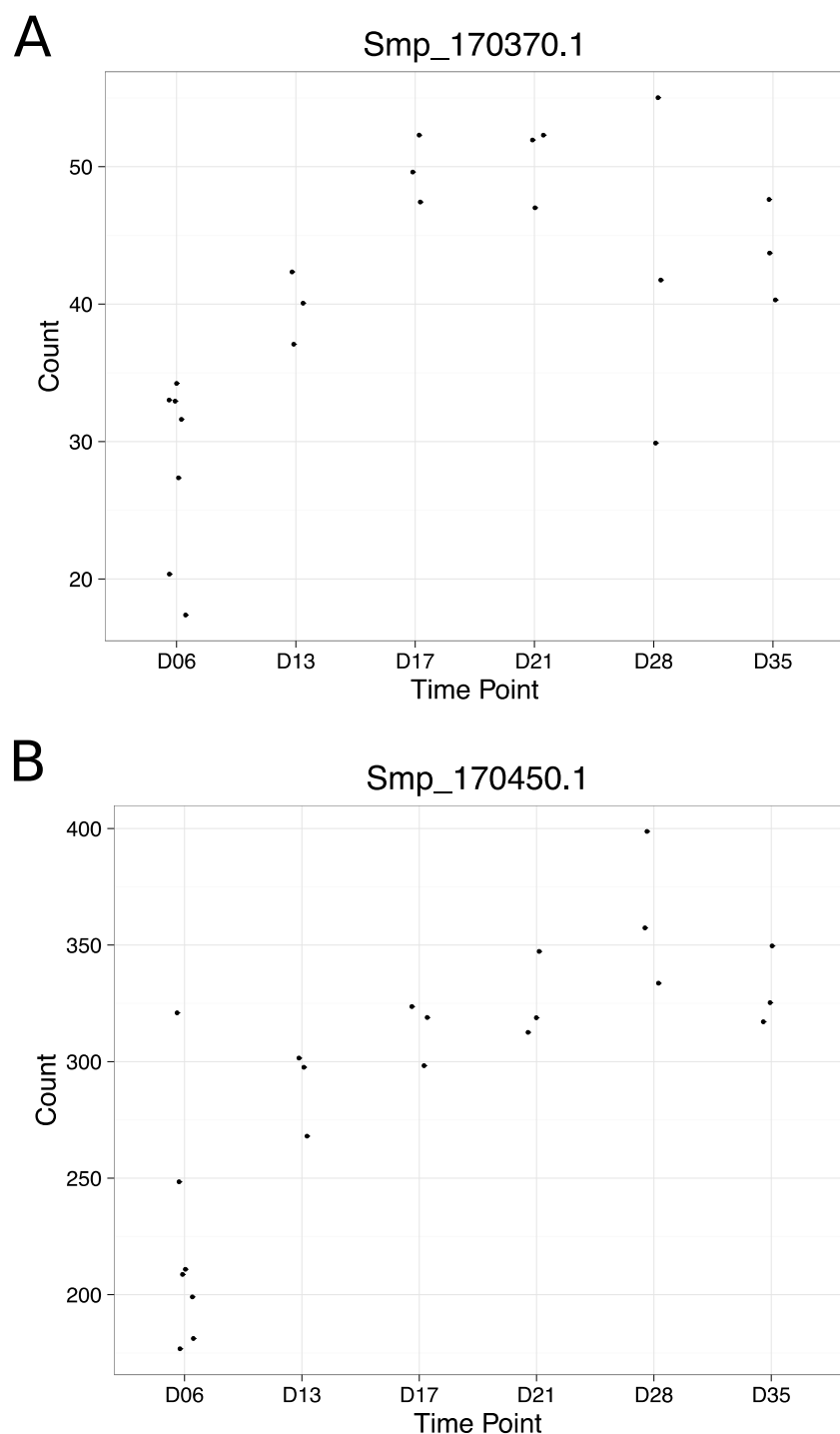


Figure 32. Expression data from RNAseq of juvenile (schistosomulae) and adult *S. mansoni* worms from an *in vivo* model from different time points (in days) post-infection.

Count refers to expression data normalised using a size factor method in DESeq2. Each point represents a biological replicate of a pool of parasites collected from a single mouse. Panel A shows expression from calcium/sodium channel gene Smp_170370 (g4127) and Panel B shows expression from calcium/ sodium channel gene Smp_170450 (g4219), see Table 11 for details.

Discussion

In this Chapter, I have demonstrated that it is possible to generate whole genome sequences from *S. mansoni* miracidia collected in the field. These data have been used to understand the population structure of *S. mansoni* in an epidemiologically important setting and scan the genome for signals of selection. Two separate scans (F_{ST} and iHS) highlighted a particular region of chromosome 2 to be under positive selection, suggestive of a hard sweep in one region (Mayuge district) in which a *de novo* mutation has conferred a selective advantage and carried proximal variants in linkage. Variants in this region were found to have the strongest association with the clearance phenotype compared against all other variants in windows of selection, suggesting a role for genes in this region driving the reduced efficacy of praziquantel observed in Chapter 3. Interestingly, this region contains two genes annotated as calcium channels, which is broadly consistent with the proposed mechanism of action of praziquantel [78, 79, 86]. Furthermore, the genes encoding these channels show increased expression in the adult worm stage, evidenced by independent RNA-seq comparing adult and schistosomula stages, and in turn is also consistent with our knowledge of the effectiveness of praziquantel against different life stages, i.e., it is active against adults but with limited effect on larval stages of *S. mansoni*. I have identified non-synonymous mutations and indels in this region, of which one variant has alleles associated with the phenotype. This provides valuable evidence for future functional and population studies to further characterise this region, confirm the causality of variants and document signals of selection in other field populations of *S. mansoni* under selective pressure from repeated rounds of MDA.

Sequencing over 200 *S. mansoni* whole genomes represented a technical challenge given the low amount of starting DNA (below the detectable level of Qubit®; 0.5 ng/ml), potential degradation in field conditions (high heat and humidity) and contamination with bacterial species in the filtered stool sample. I developed a novel protocol by comparing and optimising different methods; the best protocol involved an extraction step to elute parasite DNA from the FTA® card before ‘purifying’ the DNA through a spin column, then concentrating the elute under a heated vacuum before placing the DNA elute into a whole genome amplification reaction. A significant predictor of the proportion of sequenced reads mapping to the *S. mansoni* genome was the choice of amplification kit, which had an effect greater than the location the miracidia were picked from

(Uganda or Tanzania), the method of storage (FTA® card or buffer) and treatment of miracidia in the field (washed in nuclease free water, washed in salt buffer or washed in antibiotics). Nevertheless, despite these innovations the coverage of all but one sample fell below the expected 30x and future optimisation and evaluation studies are warranted. One plausible future methodology could involve eluting DNA from the card and then amplifying directly, further improving the quality and cost-effectiveness of the technique.

The parasite population was found to be essentially unstructured at the level of individual hosts or school as shown by a number of methods (PCA, IBD, F_{ST}). Though at larger spatial scales, such as between districts or countries, we observe structuring of the *S. mansoni* population. This lack of structure at the local level has previously been noted in Lake Victoria populations in microsatellite studies [101], and may be explained, at least in part, by the high mobility of the human definitive host in this region. The families of the infected children often practise fishing and so migrate around the shores of Lake Victoria and to the islands in the lake throughout the year to access the best waters for fishing. As hosts move around, the eggs and miracidia they shed will therefore infect local snails in different regions of the lake and these snails will go on to infect humans in these locations. This creates very high rates of migration between parasite populations. This has also been observed in other species of parasitic flatworms; the population structure of trematodes that infect salmon was influenced by the mobility of the definitive host. Parasites with a definitive host that migrates between subpopulations, such as birds, had low levels of F_{ST} [111]. Therefore the low levels of F_{ST} we observe in the *S. mansoni* miracidia may be driven by human mobility.

Analysis of the relatedness of the miracidia through IBD values appears to suggest we have at most three half-sibling parasites in the dataset; this is perhaps surprising and indicates that the 1-7 miracidia we are sampling per host, per time point does not fully account for the within-host diversity of *S. mansoni* in these foci. To capture within-host variability, denser sampling will need to be undertaken. During fieldwork, the team was able to isolate more than 100 miracidia from certain highly infective individuals in a short space of time. Therefore future work focusing on intra-individual dynamics may be able to shine a greater light on this by sequencing more miracidia from individual individuals.

By comparing miracidia from before and after treatment within the same hosts, I aimed to understand if a single round of treatment with praziquantel had an effect on the variability of these populations, with the caveat that the population we sampled from was not praziquantel naïve. Genome-wide values of nucleotide diversity suggested that there was no significant reduction in diversity following a single round of treatment at the level of whole chromosomes. A genome-wide association study also failed to find any SNPs significantly associated with time point (before or after treatment in poor clearing individuals). Differentiation along the genome of samples before and after treatment, as measured by F_{ST} was low and no region was found with an F_{ST} value greater than 0.1 (Figure 28) indicating that there are no regions of the genome showing 'great' differentiation [229].

Scans for selection along the genome, by calculating F_{ST} between parasites from different hosts or a haplotype-based method (iHS) found regions of the genome that deviate from neutral expectations (Figures 29 and 30). Of note is the fact that a narrow region at the beginning of chromosome 2 had the highest values in both tests. A test for association of individual variants with the parasite clearance phenotype of the top 1% of binned SNPs from both tests showed four SNPs to be significant after a multiple testing correction ($FDR < 0.05$), three of these variants (and those with the lowest p -values) were found in the 256 kb region of chromosome 2.

The fourth significant SNP in chromosome 4 was found to be in close proximity to genes whose only annotation was for a *Leishmania* specific glycoprotein, as *S. mansoni* is so phylogenetically distant to *Leishmania* the function of the protein may differ between the two organisms and therefore functional validation would be necessary to examine the function of these genes further.

Investigating the 256 kb region in chromosome 2 further, I found 2 calcium channels, which is of interest as the mechanism of praziquantel action has been shown to have a link with calcium channels as when their action is blocked drug efficacy is reduced [79]. Transcriptome evidence showed these genes to be expressed in the adult worm, the life stage when praziquantel acts, and the expression of these genes is increased compared with the schistosomula larval stage (against which praziquantel is ineffective in *S. mansoni* [78]). I located 4 non-synonymous SNPs at a frequency of >10% in the exons of the calcium channel genes along with 2 indels. Of these, the alleles of a SNP in predicted gene g4217 was significantly associated with the host clearance

phenotype (Table 12). While these results are in themselves insufficient to prove an association with the drug, the fact we have found a region of the genome under selection, associated with the trait of interest that are expressed in the adult worm and with a function that we would expect *a priori* is certainly interesting in the understanding of drug selection.

An important consideration in the interpretation of these data is the confounding between district and parasite clearance phenotype, as miracidia collected from Tororo district (Kocoge school) also were exclusively from 'good' clearing hosts. Examining the origin of the alternative alleles in the 256 kb region that contributing to the long haplotype identified by the iHS test, it appeared that the alternative allele was associated more with district than with time point (before or after treatment with praziquantel), therefore the power in the tests for selection appear to be driven by differences between regions with different selection pressures (different previous numbers of rounds of MDA) rather than from hosts with different 'clearance phenotypes' or different time points (before or after treatment) within areas under a strong selective pressure.

Future work should focus on the hypothetical calcium channels in the region of selection to demonstrate their phenotype. Knocking down the gene expression through RNAi or creating non-synonymous mutations through CRISPR/Cas9 and exposing the modified *S. mansoni* worms to praziquantel *in vivo* or *in vitro* would provide compelling evidence for their role in the mode of action of praziquantel [243]. Indeed, an RNAi knockdown experiment has been planned at WTSI. Although the version of the reference genome represents a significant improvement on the version used in Chapter 2, further improvements to the genome and gene models are planned in the near future meaning that the annotation around the region of selection may change. It is also possible that the functional SNP is not amongst those listed in Table 12 as it was filtered out due to it falling in a region of low depth in the samples. Due to the variable depth of the miracidia genomes, I was unable to test for structural variants (copy number variants), though these could also represent variants under selection.

Overall this Chapter represents an important step forward in understanding the population genomic structure of an *S. mansoni* population of considerable public health interest and has identified regions of the genome under selection containing calcium channels. Non-synonymous mutations within this window have been shown to be associated with the host's response to

praziquantel. If these genes are shown to regulate uptake of praziquantel, these data may play a significant role in unlocking the drug's mechanism of action, which has remained unknown since praziquantel was synthesised over 40 years ago.

Chapter 5

Conclusion and directions for future work

This thesis is the first study to generate and analyse whole-genome sequence data from populations of *Schistosoma mansoni*. These data have been used to answer questions on the demographic history of the species, including timing the migration of parasites to the New World, the population structure at global and local scales and regions of the genomes under selection. I have found evidence of positive selection in the *S. mansoni* genome at different time scales: i) since the split with the closest extant species *S. rodhaini* around 120 thousand years ago; and ii) more recently (presumably within the last decades) in a population of parasites exposed to drug pressure in Uganda. This study is also the first to make the association between the intensity of mass drug administration (MDA) and the present anthelmintic efficacy of praziquantel against *S. mansoni*. Epidemiological data on *S. mansoni* egg counts from infected children before and after treatment revealed that drug efficacy, as measured by the egg reduction rate (ERR), is correlated with past treatment history at the school level. Taken together therefore, the combination of epidemiological evidence for reduced anthelmintic efficacy in the schools most exposed to MDA and the finding that regions of the parasite genome under selection correlate significantly with the host drug response, and further that these regions contain calcium channels with non-synonymous variants, builds the evidence base that repeated MDA is affecting the parasite population and selecting variants that may confer resistance to praziquantel.

This discussion chapter aims to place these findings into a broader context and suggest areas where future research might focus.

Population history inferred from global genome sequences

My analysis in Chapter 2 on the global origins of *S. mansoni* found strong evidence that east Africa is the most likely origin of the species. I have also proposed ~120KYA as the most likely date of the split with *S. rodhaini*. By examining contemporaneous evidence from the archaeological record in Africa, I proposed that this split may be due to the changes in subsistence activity documented during this period; specifically increased fishing activity in fresh water lakes and rivers [178, 186, 244]. While previous studies have provided accounts of the origin of *S. mansoni* based on a small

number of nuclear or mitochondrial markers [20, 102, 152], the use of whole genome-sequence data and powerful coalescence models [156, 157] is an important addition to the evidence base. The study of infectious disease origins is an interesting area of research for which we are receiving more sequence data and better models, thereby giving us more power to uncover past population histories than ever before. A parallel study I have been involved with on the origins of *Plasmodium falciparum* and related parasites in great apes is also providing a much greater resolution of the population history of the *Laverania* clade using novel sequence data from a number of ape-infective *Laverania* parasites [245]. However, the estimates of dates for past events during the evolution of both *Laverania* and *Schistosoma* parasites depend on knowing the generation time and mutation rate, as these allow us to translate observed differences in genome sequence into the time taken for these differences to accumulate. More formally, the generation time and mutation rate are 'free parameters' of the coalescence models (they cannot be estimated directly from the same data, but must be ascertained independently) used to estimate the dates of speciation and ancestral population sizes.

Even in well-studied organisms, calculating the mutation rate has proved controversial as different methods can yield different results (for instance estimates from trio studies vs. comparing mutations in genomes since a known time of divergence [246]) and it is apparent that mutation rates are variable within genomes and longitudinal studies in a number of species have also shown how mutational rates have the capacity to vary through time [169, 247, 248]. As the current models for calculating the coalescence in recombining organisms are unable to account for this variability along genomes or through time, there is the potential for improved estimates of coalescence parameters in the future as methods develop.

Given the uncertainty around the dating estimates and the potential for them to shift in the future given larger datasets, improved methods or revised parameter estimates we should be careful of falling into telling 'Just So Stories' to explain our findings. The Just So Stories are children's stories by Rudyard Kipling, such as 'How the Leopard Got His Spots', in which fanciful tales are used to explain the natural world. Scientists in the field of evolutionary biology in particular have been accused of invoking untestable narratives to account for observable phenomena [249, 250]. In evolutionary genetics we should be wary of adopting these narratives, though it is impossible to

analyse historical events with the same scrutiny as contemporary phenomena and there will always be an element of speculation and inference.

Reduced efficacy of praziquantel

Mass drug administration (MDA) is set to continue as the main public health intervention against schistosomiasis in endemic settings. Efforts to increase the coverage of praziquantel should be applauded as treatment brings health benefits to populations, such as reducing anaemia and improving growth [251, 252]. In certain contexts MDA is sufficient to reduce the transmission of *S. mansoni*, typically estimated by prevalence and infection intensity [71, 73]. The global coverage of praziquantel for schistosomiasis is poor by the standard of other neglected tropical diseases (estimated at 8% of eligible people in 2009, compared with 31% for STH, 42% for lymphatic filariasis and 60% for onchocerciasis [253]), and should be improved. Valuable work by the international community, NGOs and national governments will be instrumental in this initiative, which may see the elimination of schistosomiasis within our lifetimes [254]. However, this optimism and focus on the operational side of NTD control without monitoring and evaluating mass treatment programs faces problems. Other studies have identified the biases of drug distributors at community levels for instance [255]. In Chapters 3 and 4 I highlighted how repeated rounds of treatment in a region of Uganda around Lake Victoria has failed to halt the transmission of schistosomiasis, that repeated treatment appears to be leading to reduced efficacy and that this effect is partly explained by parasite genetics indicating that resistant genes may be sweeping through the population. If such findings are found to significantly impair the treatment efficacy of praziquantel across multiple sites this would indicate a problem that should be acknowledged by the international community. Measures such as snail control or improved sanitation should be considered alongside MDA in these high intensity settings to bring about control of schistosomiasis. Measuring the treatment efficacy of praziquantel and other anthelmintics using the models implemented in Chapter 3 on a routine basis would be an excellent way of monitoring the continued efficacy of the drug and would be reassuring both for national governments trying to control the disease and for donors and pharmaceutical companies donating anthelmintics to know that treatment remains effective [62].

*Population genomics of Ugandan *S. mansoni**

In high transmission settings where parasite prevalence and infection intensity remain high despite repeated treatments, this represents a population of parasites against whom strong selection is acting and drug resistance may develop as a result. Given that we have found a potential region of the genome conferring resistance and the mutations in this region may have arisen as the result of a hard sweep, if we are able to functionally validate any of the markers highlighted in this study we will be able to monitor the frequency of the resistant variant across populations and time. Demonstrating a genetic basis for the reduced drug efficacy we observed in Chapter 3 is an important step to showing that the variance we observe are at least partly due to parasite factors, rather than solely host differences. While these developments may be concerning, a better understanding of the geographic distribution of markers for reduced praziquantel efficacy against *S. mansoni* will be valuable in shaping future control efforts. Higher doses of praziquantel or multidrug regimes, such as combining therapy with oxamniquine, could be explored in hotspots of transmission where reduced efficacy is detected.

The whole genome sequence data I generated will become more useful when the *S. mansoni* reference genome is improved to incorporate all scaffolds into chromosomes. While on-going efforts at the Wellcome Trust Sanger Institute resulted in an improved reference towards the end of my thesis in 2016, a fully completed reference would certainly improve future analysis. Completion of reference genomes has been made easier through developments in long read sequencing such as PacBio, which produces reads of several kb in length that can span the longest repeats in organisms such as *S. mansoni* and close gaps left by short read sequencing [256]. A more complete annotation of genes and their predicted function would also be a valuable resource for future population genomic studies as it becomes easier to infer which SNPs fall on which genes and hence better predict causal variants. Research in other organisms such as humans has indicated the important role of regulatory elements within the genome such as tRNAs and microRNAs in controlling gene expression and their links to disease susceptibility [257-259]; detailing the location of such elements within the *S. mansoni* genome will therefore be of considerable interest.

When full gene annotation is available for the version 6 reference genome it will be possible to apply polymorphism data such as that generated in Chapter 4 of this thesis to understand which genes are most variable in natural populations. This may indicate *S. mansoni* genes involved in

immune interactions as these genes are typically under balancing selection with higher rates of polymorphism than other coding regions [146]. Furthermore, as on-going protein screens at WTSI are attempting to identify surface-expressed proteins in *S. mansoni* that may confer protective immunity, understanding the natural variability of these proteins in parasite populations will be necessary to check the effectiveness of vaccine candidates.

While this thesis has focused on parasite genomics, it has done so at the expense of host factors such as immunology, nutritional status or host genetics. While such topics go beyond the scope of this work, the study of human health in the broadest sense makes no distinction; rather the difference is an artificial one created by academic sub-disciplines. By addressing both host and parasite factors, better science and a better understanding of disease will result, though problems will undoubtedly have to be overcome in the increased cost of sequencing hosts and pathogens alongside the challenge of integrating disparate disciplines. In doing so regions of the host and parasite genomes that co-evolve can be uncovered [260, 261].

Given the relatively high cost of sequencing large parasite genomes and the poverty of those affected by neglected tropical diseases (typically people living on <\$1 per day [262]), it is pertinent to ask what this technology brings us in understanding the disease that will bring benefits to those affected. I would argue that there is a strong case for applying this technology to understanding the genetic basis of anthelmintic efficacy, as MDA programs globally cost millions of dollars to fund [263]. Thus validation of the future success of these public health programs remains a valuable area for scientists to make a contribution. Given the size of the *S. mansoni* genome, challenges in generating high quality sequence data and the fact that much of the genome is poorly understood, a more targeted approach to sequencing could be developed such as targeting particular SNPs or exons along the genome which would allow us to answer the same important questions, though would also provide better depth of coverage and allow many more parasites to be sequenced for the same cost. The cost of sequencing is also decreasing rapidly [7], meaning that large scale sequencing projects could become more routine in the future.

Testing for selection along whole genomes is a powerful 'hypothesis-free' research method and in this study appears to have identified calcium channels that may be involved in the reduced efficacy of praziquantel. However, these results are only predictive and it is only with functional

validation that target genes may be discerned as truly causative. ‘Dry laboratory’ genomics will therefore have to work closely with ‘wet laboratory’ science to determine causative variants. The lifecycle of *S. mansoni* can be set up using a mouse model and altering the genome has been shown to be possible, using HIV1 virus or RNAi [133, 243, 264]. Advances to genome editing through CRISPR-Cas9 is revolutionising many areas of biomedical science [265] and the application of this method to *S. mansoni* would allow the function of many genes that are currently unknown to be better understood. I thereby hope this thesis will stimulate work on schistosome functional genomics as well as applying NGS approaches to understand the population structure of schistosomes in other public health relevant settings.

This thesis represents a number of ‘firsts’, the first whole genome studies of global samples of *S. mansoni* and of miracidia direct from patients. Undoubtedly, future projects will supersede this work in terms of numbers of samples, depth of sequencing and methods of analysis as they become available. However, I hope that my analysis offers a valuable ‘first draft’ of examining *S. mansoni* populations at the resolution of whole genomes and a point of comparison for future work.

References

1. Darwin C. The origin of species by means of natural selection, or, The preservation of favored races in the struggle for life. Vol. 1. International Science Library **1897**.
2. Watson JD, Crick FH. Molecular structure of nucleic acids. *Nature* **1953**; 171(4356): 737-8.
3. Huxley J. Evolution. The Modern Synthesis. *Evolution The Modern Synthesis* **1942**.
4. Nelder JA, Baker RJ. Generalized linear models. *Encyclopedia of Statistical Sciences* **1972**.
5. Dyson G. Turing's cathedral: the origins of the digital universe: Vintage, **2012**.
6. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **1977**; 74(12): 5463-7.
7. Olson PD, Hughes J, Cotton JA. Next Generation Systematics: Cambridge University Press, **2016**.
8. Fisher RA. The genetical theory of natural selection: a complete variorum edition: Oxford University Press, **1930**.
9. Lee PM. Bayesian statistics: an introduction: John Wiley & Sons, **2012**.
10. Bolker BM, Brooks ME, Clark CJ, et al. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol* **2009**; 24(3): 127-35.
11. Basch PF. Schistosomes: Development, reproduction, and host relations, **1991**.
12. Yamaguti S. Synopsis of digenetic trematodes of vertebrates. Vols I and II. Synopsis of digenetic trematodes of vertebrates Vols I and II **1971**.
13. Basch P. Why do schistosomes have separate sexes? *Parasitol Today* **1990**; 6(5): 160-3.
14. Olson PD, Tkach VV. Advances and trends in the molecular systematics of the parasitic Platyhelminthes. *Adv Parasitol* **2005**; 60: 165-243.
15. Colley DG, Bustinduy AL, Secor WE, King CH. Human schistosomiasis. *The Lancet* **2014**; 383(9936): 2253-64.
16. Boissier J, Grech-Angelini S, Webster BL, et al. Outbreak of urogenital schistosomiasis in Corsica (France): an epidemiological case study. *The Lancet Infectious Diseases* **2016**.
17. WHO. Current estimated total number of individuals with morbidity and mortality due to *Schistosoma haematobium* and *S. mansoni* infection in Sub-Saharan Africa. Available at: <http://www.who.int/schistosomiasis/epidemiology/table/en/>.
18. WHO. Yemen poised to eliminate two debilitating neglected tropical diseases in 2015.

19. Chitsulo L, Engels D, Montresor A, Savioli L. The global status of schistosomiasis and its control. *Acta Trop* **2000**; 77(1): 41-51.
20. Morgan JA, Dejong RJ, Adeoye GO, et al. Origin and diversification of the human parasite *Schistosoma mansoni*. *Mol Ecol* **2005**; 14(12): 3889-902.
21. Kabatereine NB, Brooker S, Tukahebwa EM, Kazibwe F, Onapa AW. Epidemiology and geography of *Schistosoma mansoni* in Uganda: implications for planning control. *Trop Med Int Health* **2004**; 9(3): 372-80.
22. Steinmann P, Keiser J, Bos R, Tanner M, Utzinger J. Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. *The Lancet infectious diseases* **2006**; 6(7): 411-25.
23. Anderson RM, May RM. *Infectious diseases of humans*. Oxford: Oxford University Press, **1991**.
24. Bayne CJ. Successful parasitism of vector snail *Biomphalaria glabrata* by the human blood fluke (trematode) *Schistosoma mansoni*: a 2009 assessment. *Mol Biochem Parasitol* **2009**; 165(1): 8-18.
25. Walker AJ. Insights into the functional biology of schistosomes. *Parasit Vectors* **2011**; 4(1): 1.
26. Théron A. Dynamics of larval populations of *Schistosoma mansoni* in *Biomphalaria glabrata*. II. Chronobiology of the intramolluscal larval development during the shedding period. *Ann Trop Med Parasitol* **1981**; 75(5): 547-54.
27. McKerrow J, Salter J. Invasion of skin by *Schistosoma cercariae*. *Trends Parasitol* **2002**; 18(5): 193-5.
28. Wilson R. The saga of schistosome migration and attrition. *Parasitology* **2009**; 136(12): 1581-92.
29. Lainson R, Shaw J, Cox F, Kreier J, Wakelin D. *Topley & Wilson's microbiology and microbial infections, parasitology*. Topley & Wilson's microbiology and microbial infections, parasitology **2005**.
30. Popiel I, Cioli D, Erasmus DA. The morphology and reproductive status of female *Schistosoma mansoni* following separation from male worms. *Int J Parasitol* **1984**; 14(2): 183-90.
31. Ross AG, Vickers D, Olds GR, Shah SM, McManus DP. Katayama syndrome. *The Lancet infectious diseases* **2007**; 7(3): 218-24.

32. Kolářová L, Skirnisson K, Horak P. Schistosome cercariae as the causative agent of swimmer's itch in Iceland. *J Helminthol* **1999**; 73(03): 215-20.
33. Symmers W. A new form of cirrhosis due to the ova of *Schistosoma haematobia*. *J Path Bact* **1904**; 9: 327.
34. Warren KS. The pathology, pathobiology and pathogenesis of schistosomiasis. *Nature* **1978**; 273(5664): 609-12.
35. Burke M, Jones M, Gobert G, Li Y, Ellis M, McManus D. Immunopathogenesis of human schistosomiasis. *Parasite Immunol* **2009**; 31(4): 163-76.
36. Vennervald BJ, Dunne DW. Morbidity in schistosomiasis: an update. *Curr Opin Infect Dis* **2004**; 17(5): 439-47.
37. Wolfe ND, Dunavan CP, Diamond J. Origins of major human infectious diseases. *Nature* **2007**; 447(7142): 279-83.
38. Anderson RM, May RM. Population biology of infectious diseases: Part I. **1979**.
39. Taleb NN. The black swan: The impact of the highly improbable: Random House, **2007**.
40. Cano J, Rebollo MP, Golding N, et al. The global distribution and transmission limits of lymphatic filariasis: past and present. *Parasit Vectors* **2014**; 7(1): 466.
41. Brooker S, Hotez PJ, Bundy DA. The global atlas of helminth infection: mapping the way forward in neglected tropical disease control. *PLoS Negl Trop Dis* **2010**; 4(7): e779.
42. Hay SI, Guerra CA, Gething PW, et al. A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Med* **2009**; 6(3): e1000048.
43. Gething PW, Elyazar IR, Moyes CL, et al. A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. *PLoS Negl Trop Dis* **2012**; 6(9): e1814.
44. Clements A, Moyeed R, Brooker S. Bayesian geostatistical prediction of the intensity of infection with *Schistosoma mansoni* in East Africa. *Parasitology* **2006**; 133: 711-9.
45. Brooker S. Spatial epidemiology of human schistosomiasis in Africa: risk models, transmission dynamics and control. *Trans R Soc Trop Med Hyg* **2007**; 101(1): 1-8.
46. Brooker S, Clements AC. Spatial heterogeneity of parasite co-infection: Determinants and geostatistical prediction at regional scales. *Int J Parasitol* **2009**; 39(5): 591-7.
47. Medley G, Anderson RM. Density-dependent fecundity in *Schistosoma mansoni* infections in man. *Trans R Soc Trop Med Hyg* **1985**; 79(4): 532-4.
48. Bradley DJ, May RM. Consequences of helminth aggregation for the dynamics of schistosomiasis. *Trans R Soc Trop Med Hyg* **1978**; 72(3): 262-73.

49. Fleming FM, Brooker S, Geiger SM, et al. Synergistic associations between hookworm and other helminth species in a rural community in Brazil. *Trop Med Int Health* **2006**; 11(1): 56-64.
50. Fulford A, Webster M, Ouma J, Kimani G, Dunne D. Puberty and age-related changes in susceptibility to schistosome infection. *Parasitol Today* **1998**; 14(1): 23-6.
51. Kabatereine N, Vennervald B, Ouma J, et al. Adult resistance to schistosomiasis mansoni: age-dependence of reinfection remains constant in communities with diverse exposure patterns. *Parasitology* **1999**; 118(01): 101-5.
52. Katz N, CHAVES A, Pellegrino J. A simple device for quantitative stool thick-smear technique in Schistosomiasis mansoni. *Rev Inst Med Trop Sao Paulo* **1972**; 14(6): 397-400.
53. Assefa LM, Crellen T, Kepha S, et al. Diagnostic Accuracy and Cost-Effectiveness of Alternative Methods for Detection of Soil-Transmitted Helminths in a Post-Treatment Setting in Western Kenya. *PLoS Negl Trop Dis* **2014**; 8(5): e2843.
54. Kongs A, Marks G, Verle P, Van Der Stuyft P. The unreliability of the Kato - Katz technique limits its usefulness for evaluating *S. mansoni* infections. *Trop Med Int Health* **2001**; 6(3): 163-9.
55. Lamberton PH, Kabatereine NB, Oguttu DW, Fenwick A, Webster JP. Sensitivity and specificity of multiple Kato-Katz thick smears and a circulating cathodic antigen test for *Schistosoma mansoni* diagnosis pre-and post-repeated-praziquantel treatment. **2014**.
56. Glinz D, Silue KD, Knopp S, et al. Comparing diagnostic accuracy of Kato-Katz, Koga agar plate, ether-concentration, and FLOTAC for *Schistosoma mansoni* and soil-transmitted helminths. *PLoS Negl Trop Dis* **2010**; 4(7): e754.
57. Weerakoon KG, Gobert GN, Cai P, McManus DP. Advances in the Diagnosis of Human Schistosomiasis. *Clin Microbiol Rev* **2015**; 28(4): 939-67.
58. Brooker S, Kabatereine N, Gyapong J, Stothard J, Utzinger J. Rapid mapping of schistosomiasis and other neglected tropical diseases in the context of integrated control programmes in Africa. *Parasitology* **2009**; 136(13): 1707-18.
59. Ochodo EA, Gopalakrishna G, Spek B, et al. Circulating antigen tests and urine reagent strips for diagnosis of active schistosomiasis in endemic areas. *Cochrane Database Syst Rev* **2015**; 3.

60. Ochodo EA, Gopalakrishna G, Spek B, et al. Circulating antigen tests and urine reagent strips for diagnosis of active schistosomiasis in endemic areas. *The Cochrane Library* **2015**.
61. Stothard JR, Stanton MC, Bustinduy AL, et al. Diagnostics for schistosomiasis in Africa and Arabia: a review of present options in control and future needs for elimination. *Parasitology* **2014**; 141(14): 1947-61.
62. Walker M, Churcher TS, Basáñez M-G. Models for measuring anthelmintic drug efficacy for parasitologists. *Trends Parasitol* **2014**; 30(11): 528-37.
63. el Gaddal AA. The Blue Nile Health Project: a comprehensive approach to the prevention and control of water-associated diseases in irrigated schemes of the Sudan. *The Journal of tropical medicine and hygiene* **1985**; 88(2): 47-56.
64. Fenwick A, Webster JP. Schistosomiasis: challenges for control, treatment and drug resistance. *Curr Opin Infect Dis* **2006**; 19(6): 577-82.
65. Rollinson D, Knopp S, Levitz S, et al. Time to set the agenda for schistosomiasis elimination. *Acta Trop* **2013**; 128(2): 423-40.
66. Fenwick A. New initiatives against Africa's worms. *Trans R Soc Trop Med Hyg* **2006**; 100(3): 200-7.
67. Clements AC, Lwambo NJ, Blair L, et al. Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania. *Trop Med Int Health* **2006**; 11(4): 490-503.
68. Williams E, Hayes R, Smith P. Admissions to a rural hospital in the West Nile District of Uganda over a 27 year period. *The Journal of tropical medicine and hygiene* **1986**; 89(4): 193-211.
69. Kabatereine NB, Brooker S, Koukounari A, et al. Impact of a national helminth control programme on infection and morbidity in Ugandan schoolchildren. *Bull WHO* **2007**; 85(2): 91-9.
70. Organization WH. Schistosomiasis: progress report 2001-2011, strategic plan 2012-2020: World Health Organisation, **2013**. Report No.: 9241503173.
71. French MD, Churcher TS, Gambhir M, et al. Observed reductions in *Schistosoma mansoni* transmission from large-scale administration of praziquantel in Uganda: a mathematical modelling study. *PLoS Negl Trop Dis* **2010**; 4(11): e897.

72. Touré S, Zhang Y, Bosqué-Oliva E, et al. Two-year impact of single praziquantel treatment on infection in the national control programme on schistosomiasis in Burkina Faso. *Bull WHO* **2008**; 86(10): 780-7A.
73. Ouedraogo H, Drabo F, Zongo D, et al. Schistosomiasis in school-age children in Burkina Faso after a decade of preventive chemotherapy. *Bull World Health Organ* **2016**; 94(1): 37-45.
74. Gönner R, Andrews P. Praziquantel, a new broad-spectrum antischistosomal agent. *Z Parasitenkd* **1977**; 52(2): 129-50.
75. Doenhoff MJ, Cioli D, Utzinger J. Praziquantel: mechanisms of action, resistance and new derivatives for schistosomiasis. *Curr Opin Infect Dis* **2008**; 21(6): 659-67.
76. Stelma F, Talla I, Sow S, et al. Efficacy and side effects of praziquantel in an epidemic focus of *Schistosoma mansoni*. *The American journal of tropical medicine and hygiene* **1995**; 53(2): 167-70.
77. Kabatereine N, Kemijumbi J, Ouma J, et al. Efficacy and side effects of praziquantel treatment in a highly endemic *Schistosoma mansoni* focus at Lake Albert, Uganda. *Trans R Soc Trop Med Hyg* **2003**; 97(5): 599-603.
78. Cioli D, Pica-Mattoccia L. Praziquantel. *Parasitol Res* **2003**; 90(1): S3-S9.
79. Pica-Mattoccia L, Valle C, Basso A, et al. Cytochalasin D abolishes the schistosomicidal activity of praziquantel. *Exp Parasitol* **2007**; 115(4): 344-51.
80. Brindley P, Sher A. The chemotherapeutic effect of praziquantel against *Schistosoma mansoni* is dependent on host antibody response. *The Journal of Immunology* **1987**; 139(1): 215-20.
81. Fallon PG, Doenhoff MJ. Drug-resistant schistosomiasis: resistance to praziquantel and oxamniquine induced in *Schistosoma mansoni* in mice is drug specific. *Am J Trop Med Hyg* **1994**; 51(1): 83-8.
82. Assessing the efficacy of anthelmintic drugs against schistosomiasis and soil-transmitted helminthiases. *World Health Organization* **2013**.
83. Alexander N. Review: analysis of parasite and other skewed counts. *Trop Med Int Health* **2012**; 17(6): 684-93.
84. Walker M, Mabud TS, Oliaro PL, et al. New approaches to measuring anthelmintic drug efficacy: parasitological responses of childhood schistosome infections to treatment with praziquantel. *Parasit Vectors* **2016**; 9(1): 1.

85. Doenhoff MJ, Kusel JR, Coles GC, Cioli D. Resistance of *Schistosoma mansoni* to praziquantel: is there a problem? *Trans R Soc Trop Med Hyg* **2002**; 96(5): 465-9.
86. Greenberg RM. New approaches for understanding mechanisms of drug resistance in schistosomes. *Parasitology* **2013**; 140(12): 1534-46.
87. Southgate V, Tchuem Tchuente L, Sene M, et al. Studies on the biology of schistosomiasis with emphasis on the Senegal river basin. *Mem Inst Oswaldo Cruz* **2001**; 96: 75-8.
88. Danso-Appiah A, De Vlas SJ. Interpreting low praziquantel cure rates of *Schistosoma mansoni* infections in Senegal. *Trends Parasitol* **2002**; 18(3): 125-9.
89. Webster BL, Diaw OT, Seye MM, et al. Praziquantel treatment of school children from single and mixed infection foci of intestinal and urogenital schistosomiasis along the Senegal River Basin: monitoring treatment success and re-infection patterns. *Acta Trop* **2013**; 128(2): 292-302.
90. Wang W, Wang L, Liang Y-S. Susceptibility or resistance of praziquantel in human schistosomiasis: a review. *Parasitol Res* **2012**; 111(5): 1871-7.
91. Barakat R, MORSHEDY HE. Efficacy of two praziquantel treatments among primary school children in an area of high *Schistosoma mansoni* endemicity, Nile Delta, Egypt. *Parasitology* **2011**; 138(04): 440-6.
92. Crellen T, Walker M, Lamberton PH, et al. Reduced efficacy of praziquantel against *Schistosoma mansoni* is associated with multiple-rounds of mass drug administration. *Clin Infect Dis* **2016**: ciw506.
93. Darwin C, Bynum WF. The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life. **2009**.
94. Fletcher M, LoVerde PT, Woodruff DS. Genetic variation in *Schistosoma mansoni*: enzyme polymorphisms in populations from Africa, Southwest Asia, South America, and the West Indies. *Am J Trop Med Hyg* **1981**; 30: 406-21.
95. Gower CM, Gouvras AN, Lamberton PH, et al. Population genetic structure of *Schistosoma mansoni* and *Schistosoma haematobium* from across six sub-Saharan African countries: Implications for epidemiology, evolution and control. *Acta Trop* **2013**; 128(2): 261-74.
96. Hebert PD, Ratnasingham S, de Waard JR. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences* **2003**; 270(Suppl 1): S96-S9.

97. Taylor H, Harris W. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources* **2012**; 12(3): 377-88.
98. Collins R, Cruickshank R. The seven deadly sins of DNA barcoding. *Molecular ecology resources* **2013**; 13(6): 969-75.
99. Zarowiecki M, Huyse T, Littlewood D. Making the most of mitochondrial genomes—markers for phylogeny, molecular ecology and barcodes in *Schistosoma* (Platyhelminthes: Digenea). *Int J Parasitol* **2007**; 37(12): 1401-18.
100. Nichols R. Gene trees and species trees are not the same. *Trends Ecol Evol* **2001**; 16(7): 358-64.
101. Stothard J, Webster B, Weber T, et al. Molecular epidemiology of *Schistosoma mansoni* in Uganda: DNA barcoding reveals substantial genetic diversity within Lake Albert and Lake Victoria populations. *Parasitology* **2009**; 136(13): 1813-24.
102. Webster BL, Webster JP, Gouvras AN, et al. DNA 'barcoding' of *Schistosoma mansoni* across sub-Saharan Africa supports substantial within locality diversity and geographical separation of genotypes. *Acta Trop* **2013**; 128(2): 250-60.
103. Leles D, Gardner SL, Reinhard K, Iñiguez A, Araujo A. Are *Ascaris lumbricoides* and *Ascaris suum* a single species? *Parasit Vectors* **2012**; 5(1): 1.
104. Zhu X, Gasser RB, Jacobs DE, Hung G-C, Chilton NB. Relationships among some ascaridoid nematodes based on ribosomal DNA sequence data. *Parasitol Res* **2000**; 86(9): 738-44.
105. Liu G-H, Wu C-Y, Song H-Q, et al. Comparative analyses of the complete mitochondrial genomes of *Ascaris lumbricoides* and *Ascaris suum* from humans and pigs. *Gene* **2012**; 492(1): 110-6.
106. Betson M, Nejsun P, Bendall RP, Deb RM, Stothard JR. Molecular epidemiology of ascariasis: a global perspective on the transmission dynamics of *Ascaris* in people and pigs. *J Infect Dis* **2014**; 210(6): 932-41.
107. Huyse T, Webster BL, Geldof S, et al. Bidirectional introgressive hybridization between a cattle and human schistosome species. *PLoS Path* **2009**; 5(9): e1000571.
108. Morgan JA, DeJong RJ, Lwambo NJ, Mungai BN, Mkoji GM, Loker ES. First report of a natural hybrid between *Schistosoma mansoni* and *S. rodhaini*. *J Parasitol* **2003**; 89(2): 416-8.
109. LeRoux P. Hybridization of *Schistosoma mansoni* and *S. rodhaini*. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **1954**; 48(1): 3-4.

110. Stukenbrock EH, Christiansen FB, Hansen TT, Dutheil JY, Schierup MH. Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. *Proceedings of the National Academy of Sciences* **2012**; 109(27): 10954-9.
111. Criscione CD, Blouin MS. Life cycles shape parasite evolution: comparative population genetics of salmon trematodes. *Evolution* **2004**; 58(1): 198-202.
112. Prugnolle F, De Meeûs T, Durand P, Sire C, Théron A. Sex - specific genetic structure in *Schistosoma mansoni*: evolutionary and epidemiological implications. *Mol Ecol* **2002**; 11(7): 1231-8.
113. Norton AJ, Gower CM, Lamberton PH, et al. Genetic consequences of mass human chemotherapy for *Schistosoma mansoni*: population structure pre- and post-praziquantel treatment in Tanzania. *Am J Trop Med Hyg* **2010**; 83(4): 951-7.
114. Huyse T, Van den Broeck F, Jombart T, et al. Regular treatments of praziquantel do not impact on the genetic make-up of *Schistosoma mansoni* in Northern Senegal. *Infect Genet Evol* **2013**; 18: 100-5.
115. French MD, Churcher TS, Gambhir M, et al. Observed reductions in *Schistosoma mansoni* transmission from large-scale administration of praziquantel in Uganda: a mathematical modelling study. *PLoS Negl Trop Dis* **2010**; 4(11): e897.
116. Lelo AE, Mburu DN, Magoma GN, et al. No apparent reduction in schistosome burden or genetic diversity following four years of school-based mass drug administration in mwea, central kenya, a heavy transmission area. *PLoS neglected tropical diseases* **2014**; 8(10): e3221.
117. Blanton RE, Blank WA, Costa JM, et al. *Schistosoma mansoni* population structure and persistence after praziquantel treatment in two villages of Bahia, Brazil. *Int J Parasitol* **2011**; 41(10): 1093-9.
118. Jost LOU. GST and its relatives do not measure differentiation. *Mol Ecol* **2008**; 17(18): 4015-26.
119. Allendorf FW, Hohenlohe PA, Luikart G. Genomics and the future of conservation genetics. *Nat Rev Genet* **2010**; 11(10): 697-709.
120. Pampoulie C, Ruzzante DE, Chosson V, et al. The genetic structure of Atlantic cod (*Gadus morhua*) around Iceland: insight from microsatellites, the PanI locus, and tagging experiments. *Can J Fish Aquat Sci* **2006**; 63(12): 2660-74.

121. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **2016**; 17(6): 333-51.
122. Illumina. Illumina Sequencing Technology. Available at: <http://www.illumina.com/technology/next-generation-sequencing/sequencing-technology.html>. Accessed 4th August 2016.
123. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **2011**; 12(6): 443-51.
124. Protasio AV, Tsai IJ, Babbage A, et al. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS Negl Trop Dis* **2012**; 6(1): e1455.
125. Berriman M, Haas BJ, LoVerde PT, et al. The genome of the blood fluke *Schistosoma mansoni*. *Nature* **2009**; 460(7253): 352-8.
126. Tsai IJ, Zarowiecki M, Holroyd N, et al. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* **2013**; 496(7443): 57-63.
127. Consortium TSjGSaFA. The *Schistosoma japonicum* genome reveals features of host-parasite interplay. *Nature* **2009**; 460(7253): 345-51.
128. Young ND, Jex AR, Li B, et al. Whole-genome sequence of *Schistosoma haematobium*. *Nat Genet* **2012**; 44(2): 221-5.
129. Young ND, Chan K-G, Korhonen PK, et al. Exploring molecular variation in *Schistosoma japonicum* in China. *Scientific reports* **2015**; 5.
130. Chilton N, Bao-Zhen Q, Bøgh H, Nansen P. An electrophoretic comparison of *Schistosoma japonicum* (Trematoda) from different provinces in the People's Republic of China suggests the existence of cryptic species. *Parasitology* **1999**; 119(04): 375-83.
131. Chevalier FD, Valentim CL, LoVerde PT, Anderson TJ. Efficient linkage mapping using exome capture and extreme QTL in schistosome parasites. *BMC Genomics* **2014**; 15(1): 617.
132. Crellen T, Allan F, David S, et al. Whole genome resequencing of the human parasite *Schistosoma mansoni* reveals population history and effects of selection. *Scientific Reports* **2016**; 6: 20954.
133. Valentim CL, Cioli D, Chevalier FD, et al. Genetic and molecular basis of drug resistance and species-specific drug action in schistosome parasites. *Science* **2013**; 342(6164): 1385-9.

134. Manske M, Miotto O, Campino S, et al. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* **2012**; 487(7407): 375-9.
135. Phyto AP, Nkhoma S, Stepniewska K, et al. Emergence of artemisinin-resistant malaria on the western border of Thailand: a longitudinal study. *The Lancet* **2012**; 379(9830): 1960-6.
136. Dondorp AM, Nosten F, Yi P, et al. Artemisinin resistance in *Plasmodium falciparum* malaria. *New Engl J Med* **2009**; 361(5): 455-67.
137. Cheeseman IH, Miller BA, Nair S, et al. A major genome region underlying artemisinin resistance in malaria. *Science* **2012**; 336(6077): 79-82.
138. Sabeti PC, Varilly P, Fry B, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature* **2007**; 449(7164): 913-8.
139. Ariey F, Witkowski B, Amaratunga C, et al. A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature* **2014**; 505(7481): 50-5.
140. Lamberton PH, Crellen T, Cotton JA, Webster JP. Chapter Seven-Modelling the Effects of Mass Drug Administration on the Molecular Epidemiology of Schistosomes. *Adv Parasitol* **2015**; 87: 293-327.
141. Fan S, Hansen ME, Lo Y, Tishkoff SA. Going global by adapting local: A review of recent human adaptation. *Science* **2016**; 354(6308): 54-9.
142. Protasio AV, Dunne DW, Berriman M. Comparative study of transcriptome profiles of mechanical- and skin-transformed *Schistosoma mansoni* schistosomula. *PLoS Negl Trop Dis* **2013**; 7(3): e2091.
143. Schwetz J. On two schistosomes of wild rodents of the Belgian Congo: *Schistosoma rodhaini*; and *Schistosoma mansoni* var. *Rodentorum*; and their relationship to *S. mansoni* of man. *Trans R Soc Trop Med Hyg* **1954**; 48(1): 89-100.
144. Morgan J, Dejong R, Snyder S, Mkoji G, Loker E. *Schistosoma mansoni* and *Biomphalaria*: past history and future trends. *Parasitology* **2001**; 123(07): 211-28.
145. Otto TD, Rayner JC, Böhme U, et al. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat Commun* **2014**; 5.
146. Conway DJ. Paths to a malaria vaccine illuminated by parasite genomics. *Trends Genet* **2015**; 31(2): 97-107.
147. Cornejo OE, Fisher D, Escalante AA. Genome-Wide Patterns of Genetic Polymorphism and Signatures of Selection in *Plasmodium vivax*. *Genome Biol Evol* **2014**; 7(1): 1.

148. Sironi M, Cagliani R, Forni D, Clerici M. Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nature Reviews Genetics* **2015**; 16(4): 224-36.
149. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. *Nature Reviews Genetics* **2014**; 15(6): 379-93.
150. Despres L, Imbert-Establet D, Combes C, Bonhomme F. Molecular evidence linking hominid evolution to recent radiation of schistosomes (Platyhelminthes: Trematoda). *Mol Phylogen Evol* **1992**; 1(4): 295-304.
151. Hubendick B. Studies on *Bulinus*. *Ark. Zool.* 40A, **1948**.
152. Després L, Imbert-Establet D, Monnerot M. Molecular characterization of mitochondrial DNA provides evidence for the recent introduction of *Schistosoma mansoni* into America. *Mol Biochem Parasitol* **1993**; 60(2): 221-9.
153. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **2014**; 46(8): 919-25.
154. Rosenberg NA, Nordborg M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics* **2002**; 3(5): 380-90.
155. Pritchard JK. Whole-genome sequencing data offer insights into human demography. *Nat Genet* **2011**; 43(10): 923-5.
156. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature* **2011**; 475(7357): 493-6.
157. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet* **2011**; 43(10): 1031-4.
158. Xue Y, Prado-Martinez J, Sudmant PH, et al. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* **2015**; 348(6231): 242-5.
159. Emery AM, Allan FE, Rabone ME, Rollinson D. Schistosomiasis collection at NHM (SCAN). *Parasit Vectors* **2012**; 5: 185.
160. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **2011**; 43(5): 491-8.
161. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**; 25(16): 2078-9.
162. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences: CABIOS* **1997**; 13(5): 555-6.

163. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* **2013**; 8(8): 1551-66.
164. Otto TD, Rayner JC, Bohme U, et al. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat Commun* **2014**; 5(4754): 4754.
165. Holloway AK, Lawniczak MK, Mezey JG, Begun DJ, Jones CD. Adaptive gene expression divergence inferred from population genomics. *PLoS Genet* **2007**; 3(10): e187.
166. Delaneau O, Marchini J, Consortium GP. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun* **2014**; 5.
167. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**; 30(9): 1312-3.
168. Rambaut A. FigTree, a graphical viewer of phylogenetic trees. Available at: <http://treebioedacuk/software/figtree> (Accessed: February 2015) **2007**.
169. Lynch M. Evolution of the mutation rate. *Trends Genet* **2010**; 26(8): 345-52.
170. Faust EC, Hoffman WA. Studies on schistosomiasis mansoni in Puerto Rico. III. Biological studies. 1. The extra mammalian phases of the life cycle. *Puerto Rico Journal of Public Health and Tropical Medicine* **1934**; 10: 1-47.
171. Crellen T, Iantorno S. A switch in time. *Nature Reviews Microbiology* **2015**; 13: 190-1.
172. Farias LP, Rodrigues D, Cunha V, Rofatto HK, Faquim-Mauro EL, Leite LC. Schistosoma mansoni venom allergen like proteins present differential allergic responses in a murine model of airway inflammation. *PLoS Negl Trop Dis* **2012**; 6(2): e1510.
173. Salter JP, Choe Y, Albrecht H, et al. Cercarial elastase is encoded by a functionally conserved gene family across multiple species of schistosomes. *J Biol Chem* **2002**; 277(27): 24618-24.
174. Studer RA, Robinson-Rechavi M. Large-scale Analyses of Positive Selection Using Codon Models. In: Pontarotti P. *Evolutionary Biology: Concept, Modeling, and Application*: Springer Berlin Heidelberg, **2009**:217-35.
175. Gharib WH, Robinson-Rechavi M. The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Mol Biol Evol* **2013**; 30(7): 1675-86.
176. Steinauer ML, Hanelt B, Mwangi IN, et al. Introgressive hybridization of human and rodent schistosome parasites in western Kenya. *Mol Ecol* **2008**; 17(23): 5062-74.

177. Geggus D. The French slave trade: an overview. *The William and Mary Quarterly* **2001**; 58(1): 119-38.
178. Yellen JE, Brooks AS, Cornelissen E, Mehlman MJ, Stewart K. A middle stone age worked bone industry from Katanda, Upper Semliki Valley, Zaire. *Science* **1995**; 268(5210): 553-6.
179. Walter RC, Buffler RT, Bruggemann JH, et al. Early human occupation of the Red Sea coast of Eritrea during the last interglacial. *Nature* **2000**; 405(6782): 65-9.
180. Marean CW, Bar-Matthews M, Bernatchez J, et al. Early human use of marine resources and pigment in South Africa during the Middle Pleistocene. *Nature* **2007**; 449(7164): 905-8.
181. Bos KI, Harkins KM, Herbig A, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **2014**; 514(7523): 494-7.
182. Harper KN, Zuckerman MK, Harper ML, Kingston JD, Armelagos GJ. The origin and antiquity of syphilis revisited: An Appraisal of Old World pre - Columbian evidence for treponemal infection. *Am J Phys Anthropol* **2011**; 146(S53): 99-133.
183. Leblois R, Kuhls K, François O, Schönián G, Wirth T. Guns, germs and dogs: On the origin of *Leishmania chagasi*. *Infect, Genet Evol* **2011**; 11(5): 1091-5.
184. McDougall I, Brown FH, Fleagle JG. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* **2005**; 433(7027): 733-6.
185. Clark JD. The Middle Stone Age of East Africa and the beginnings of regional identity. *Journal of World Prehistory* **1988**; 2(3): 235-305.
186. McBrearty S. Down with the revolution. In: Mellars P, Boyle K, Bar-Yosef O, Stringer C. *Rethinking the human revolution*. Cambridge: MacDonald Institute for Archaeological Research Monographs, **2007**:133-52.
187. Brooks AS, Helgren DM, Cramer JS, et al. Dating and context of three Middle Stone Age sites with bone points in the Upper Semliki Valley, Zaire. *Science* **1995**; 268(5210): 548-53.
188. Tierney JE, Russell JM, Huang Y, Damsté JSS, Hopmans EC, Cohen AS. Northern hemisphere controls on tropical southeast African climate during the past 60,000 years. *Science* **2008**; 322(5899): 252-5.
189. Sturmbauer C, Baric S, Salzburger W, Rüber L, Verheyen E. Lake level fluctuations synchronize genetic divergences of cichlid fishes in African lakes. *Mol Biol Evol* **2001**; 18(2): 144-54.

190. Campbell MC, Tishkoff SA. The evolution of human genetic and phenotypic variation in Africa. *Curr Biol* **2010**; 20(4): R166-R73.
191. Anderson R, May R. Prevalence of schistosome infections within molluscan populations: observed patterns and theoretical predictions. *Parasitology* **1979**; 79(01): 63-94.
192. Lynch M. Evolution of the mutation rate. *Trends Genet* **2010**; 26(8): 345-52.
193. Halligan DL, Keightley PD. Spontaneous mutation accumulation studies in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics* **2009**; 40: 151-72.
194. Wallis JPR, Livingstone D, Hiller V. The Zambezi Expedition of David Livingstone 1858-1863. London: Chatto & Windus, **1956**.
195. Finn RD, Bateman A, Clements J, et al. Pfam: the protein families database. *Nucleic Acids Res* **2014**; 42(D1): D222-D30.
196. Tsai IJ, Zarowiecki M, Holroyd N, et al. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* **2013**; 496(7443): 57-63.
197. Hoffmann KF, Brindley PJ, Berriman M. Halting harmful helminths. *Science* **2014**; 346(6206): 168-9.
198. Fenwick A, Webster JP, Bosque-Oliva E, et al. The Schistosomiasis Control Initiative (SCI): rationale, development and implementation from 2002–2008. *Parasitology* **2009**; 136(13): 1719-30.
199. Fleming FM, Harrison W, Fenwick A. Schistosomiasis Control Initiative Website. Accessed 13th January.
200. Knopp S, Stothard JR, Rollinson D, et al. From morbidity control to transmission control: time to change tactics against helminths on Unguja Island, Zanzibar. *Acta Trop* **2013**; 128(2): 412-22.
201. Lahiri SN. Resampling methods for dependent data: Springer Science & Business Media, **2013**.
202. Elston D, Moss R, Boulinier T, Arrowsmith C, Lambin X. Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks. *Parasitology* **2001**; 122(05): 563-9.
203. Hadfield JD. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software* **2010**; 33(2): 1-22.
204. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015. URL <http://www.R-project.org> **2016**.
205. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis: Taylor & Francis, **2014**.

206. Mutapi F, Ndhlovu PD, Hagan P, et al. Chemotherapy accelerates the development of acquired immune responses to *Schistosoma haematobium* infection. *J Infect Dis* **1998**; 178(1): 289-93.
207. Zwang J, Olliaro PL. Clinical efficacy and tolerability of praziquantel for intestinal and urinary schistosomiasis—a meta-analysis of comparative and non-comparative clinical trials. *PLoS Negl Trop Dis* **2014**; 8(11): e3286.
208. Wolstenholme AJ, Fairweather I, Prichard R, von Samson-Himmelstjerna G, Sangster NC. Drug resistance in veterinary helminths. *Trends Parasitol* **2004**; 20(10): 469-76.
209. Bourguinat C, Pion SD, Kamgno J, et al. Genetic selection of low fertile *Onchocerca volvulus* by ivermectin treatment. *PLoS Negl Trop Dis* **2007**; 1(1): e72.
210. Webster JP, Molyneux DH, Hotez PJ, Fenwick A. The contribution of mass drug administration to global health: past, present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2014**; 369(1645): 20130434.
211. Aragon AD, Imani RA, Blackburn VR, et al. Towards an understanding of the mechanism of action of praziquantel. *Mol Biochem Parasitol* **2009**; 164(1): 57-65.
212. Lees JA, Vehkala M, Välimäki N, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *bioRxiv* **2016**: 038463.
213. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. *Annu Rev Genet* **2013**; 47: 97-120.
214. Chan JD, Zarowiecki M, Marchant JS. Ca 2+ channels and praziquantel: a view from the free world. *Parasitol Int* **2013**; 62(6): 619-28.
215. Gower CM, Gouvras AN, Lamberton PH, et al. Population genetic structure of *Schistosoma mansoni* and *Schistosoma haematobium* from across six sub-Saharan African countries: implications for epidemiology, evolution and control. *Acta Trop* **2013**; 128(2): 261-74.
216. Charlesworth B. *Elements of evolutionary genetics*: Roberts Publishers, **2010**.
217. Lelo AE, Mburu DN, Magoma GN, et al. No apparent reduction in schistosome burden or genetic diversity following four years of school-based mass drug administration in Mwea, central Kenya, a heavy transmission area. *PLoS Negl Trop Dis* **2014**; 8(10): e3221.
218. Cheever AW. A quantitative post-mortem study of schistosomiasis mansoni in man. *The American journal of tropical medicine and hygiene* **1968**; 17(1): 38-64.
219. Dozmorov MG, Adrianto I, Giles CB, et al. Detrimental effects of duplicate reads and low complexity regions on RNA-and ChIP-seq data. *BMC Bioinformatics* **2015**; 16(13): 1.

220. Pitchford R, Visser P. A simple and rapid technique for quantitative estimation of helminth eggs in human and animal excreta with special reference to *Schistosoma* sp. *Trans R Soc Trop Med Hyg* **1975**; 69(3): 318-22.
221. Gower C, Shrivastava J, Lamberton P, et al. Development and application of an ethically and epidemiologically advantageous assay for the multi-locus microsatellite analysis of *Schistosoma mansoni*. *Parasitology* **2007**; 134(04): 523-36.
222. Sigma-Aldrich. Whatman FTA card technology. Available at: <http://www.sigmaaldrich.com/catalog/product/sigma/z719730?lang=en®ion=GB>. Accessed 5th December.
223. Lamble S, Batty E, Attar M, et al. Improved workflows for high throughput library preparation using the transposome-based nextera system. *BMC Biotechnol* **2013**; 13(1): 1.
224. Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high - confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics* **2013**; 11.0. 1-.0. 33.
225. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **2015**; 4(1): 1.
226. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. *Bioinformatics* **2010**; 26(22): 2867-73.
227. Whitlock MC. Fixation probability and time in subdivided populations. *Genetics* **2003**; 164(2): 767-79.
228. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST} . *Nature Reviews Genetics* **2009**; 10(9): 639-50.
229. Hartl DL, Clark AG, Clark AG. *Principles of population genetics*: Sinauer associates Sunderland, **1997**.
230. Wickham H. *ggplot2: elegant graphics for data analysis*: Springer Science & Business Media, **2009**.
231. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc* **2010**; 5(9): 1564-73.
232. Manske M, Miotto O, Campino S, et al. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* **2012**; 487(7407): 375-9.

233. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics* **2014**; 15(5): 335-46.
234. Mackay TF, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **2009**; 10(8): 565-77.
235. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS Biol* **2006**; 4(3): e72.
236. Criscione CD, Valentim CL, Hirai H, LoVerde PT, Anderson TJ. Genomic linkage map of the human blood fluke *Schistosoma mansoni*. *Genome Biol* **2009**; 10(6): 1.
237. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **2012**; 9(2): 179-81.
238. Szpiech ZA, Hernandez RD. Selscan: an efficient multi-threaded program to perform EHH-based scans for positive selection. *Mol Biol Evol* **2014**: msu211.
239. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **2012**; 6(2): 80-92.
240. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution* **1984**: 1358-70.
241. Noble WS. How does multiple testing correction work? *Nat Biotechnol* **2009**; 27(12): 1135-7.
242. Bouvier J, Bordier C, Vogel H, Reichelt R, Etges R. Characterization of the promastigote surface protease of *Leishmania* as a membrane-bound zinc endopeptidase. *Mol Biochem Parasitol* **1989**; 37(2): 235-45.
243. Suttiaprapa S, Rinaldi G, Tsai IJ, et al. HIV-1 Integrates Widely throughout the Genome of the Human Blood Fluke *Schistosoma mansoni*. *PLoS Pathog* **2016**; 12(10): e1005931.
244. Henshilwood CS, d'Errico F, Marean CW, Milo RG, Yates R. An early bone tool industry from the Middle Stone Age at Blombos Cave, South Africa: implications for the origins of modern human behaviour, symbolism and language. *J Hum Evol* **2001**; 41(6): 631-78.
245. Otto TD, Gilabert A, Crellen T, et al. Genomes of an entire *Plasmodium* subgenus reveal paths to virulent human malaria. *bioRxiv* **2016**: 095679.
246. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics* **2012**; 13(10): 745-53.

247. Cui Y, Yu C, Yan Y, et al. Historical variations in mutation rate in an epidemic pathogen, *Yersinia pestis*. *Proceedings of the National Academy of Sciences* **2013**; 110(2): 577-82.
248. Ségurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* **2014**; 15: 47-70.
249. Gould SJ, Lewontin RC. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London B: Biological Sciences* **1979**; 205(1161): 581-98.
250. Pigliucci M, Kaplan J. The fall and rise of Dr Pangloss: adaptationism and the Spandrels paper 20 years later. *Trends Ecol Evol* **2000**; 15(2): 66-70.
251. Koukounari A, Fenwick A, Whawell S, et al. Morbidity indicators of *Schistosoma mansoni*: relationship between infection and anemia in Ugandan schoolchildren before and after praziquantel and albendazole chemotherapy. *The American journal of tropical medicine and hygiene* **2006**; 75(2): 278-86.
252. Gurarie D, Wang X, Bustinduy AL, King CH. Modeling the effect of chronic schistosomiasis on childhood development and the potential for catch-up growth with different drug treatment strategies promoted for control of endemic schistosomiasis. *Am J Trop Med Hyg* **2011**; 84(5): 773-81.
253. Organization WH. Accelerating work to overcome the global impact of neglected tropical diseases: a roadmap for implementation: executive summary. **2012**.
254. Savioli L, Fenwick A, Rollinson D, Albonico M, Ame SM. An achievable goal: control and elimination of schistosomiasis. *The Lancet* **2015**; 386(9995): 739.
255. Chami GF, Kontoleon AA, Bulte E, et al. Profiling Nonrecipients of Mass Drug Administration for Schistosomiasis and Hookworm Infections: A Comprehensive Analysis of Praziquantel and Albendazole Coverage in Community-Directed Treatment in Uganda. *Clin Infect Dis* **2015**: civ829.
256. English AC, Richards S, Han Y, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **2012**; 7(11): e47768.
257. Consortium EP, Bernstein BE, Birney E, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**; 489(7414): 57-74.
258. Gerstein MB, Kundaje A, Hariharan M, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **2012**; 489(7414): 91-100.

- 259. Grundberg E, Meduri E, Sandling JK, et al. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *The American Journal of Human Genetics* **2013**; 93(5): 876-90.
- 260. Bartha I, Carlson JM, Brumme CJ, et al. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife* **2013**; 2: e01123.
- 261. Woolhouse M, Webster J. In search of the Red Queen. *Parasitol Today* **2000**; 16(12): 506-8.
- 262. Hotez PJ, Kamath A. Neglected tropical diseases in sub-saharan Africa: review of their prevalence, distribution, and disease burden. *PLoS Negl Trop Dis* **2009**; 3(8): e412.
- 263. Webster JP, Molyneux DH, Hotez PJ, Fenwick A. The contribution of mass drug administration to global health: past, present and future. *Phil Trans R Soc B* **2014**; 369(1645): 20130434.
- 264. Patocka N, Ribeiro P. The functional role of a serotonin transporter in *Schistosoma mansoni* elucidated through immunolocalization and RNA interference (RNAi). *Mol Biochem Parasitol* **2013**; 187(1): 32-42.
- 265. Doudna JA, Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. *Science* **2014**; 346(6213): 1258096.

Appendix 1

Full laboratory protocol used to generate whole-genome sequence data from *S. mansoni* miracidia collected in Uganda (Chapter 4)

DNA Extraction and purification from FTA® Card (using buffers from Qiagen DNeasy Blood and Tissue Extraction Kit)

- 1) Miracidium DNA was punched from FTA® card using Harris Micropunch into 1.5ml microcentrifuge tube, 45µl ATL buffer and 5µl proteinase K were added to sample before vortexing and incubating at 56°C on a rocking platform for 1 hour
- 2) Add 50µl of Buffer AL and vortex, then 50µl of 100% ethanol and mix again by vortexing
- 3) Pipette out all 150µl liquid, squashing liquid out of the punch with the pipette tip and apply mixture to a spin column
- 4) Centrifuge at 8000rpm for 1 min, discard collection tube, add new collection tube
- 5) Add 500µl Buffer AW1 to spin column and centrifuge for 1min at 8000rpm for 1 min, discard collection tube and add new collection tube
- 6) Add 500µl Buffer AW2 to spin column and centrifuge for 3 mins at 14,000rpm. Discard collection tube and add 1.5ml microcentrifuge tube to the bottom of the spin column
- 7) Pipette 50µl of Buffer AE heated to 30-37°C directly on the spin column membrane and centrifuge at 1 min at 8000 rpm to elute DNA; repeat once.

DNA amplification (using reagents from GenomiPhi V3 amplification kit):

- 8) The DNA elute was then concentrated and used in an amplification reaction. Dry down samples to 10-50µl using Savant SpeedVac at 30°C, being careful not to 'over-dry' the samples as this may cause damage to the DNA.
- 9) 10ul of 2x denaturation buffer was added to the sample, which was then heated to 95°C for 3 minutes and cooled to 4°C on ice.
- 10) The denatured product was added to GenomiPhi freeze dried 'cake' and incubated at 30°C for 90 minutes. Then cooled to 4°C on ice.

DNA clean up and library preparation

- 11) Samples were purified with magnetic bead DNA cleanup and eluted into 50µls for library preparation.
- 12) Illumina libraries were produced by Sanger Core facilities for HiSeqX10 platform-CustomWGS-AUTO libraries with at least 8 PCR cycles. Where the library concentration was low, additional PCR cycles were applied.

Appendix 2

First page from Lamberton *et al.* (2015) publication in *Advances in Parasitology*. Full text available at <http://www.sciencedirect.com/science/article/pii/S0065308X14000074>

Modelling the Effects of Mass Drug Administration on the Molecular Epidemiology of Schistosomes

Poppy H.L. Lamberton^{*}, Thomas Crellen^{*,§}, James A. Cotton[§],
Joanne P. Webster^{*,¶,1}

^{*}Department of Infectious Disease Epidemiology, School of Public Health, Faculty of Medicine, Imperial College London, St Mary's Campus, London, UK

[§]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK

[¶]Department of Pathology and Pathogen Biology, Centre for Emerging, Endemic and Exotic Diseases (CEEED), Royal Veterinary College, University of London, London, UK

¹Corresponding author: E-mail: jowebster@rvc.ac.uk

Contents

1. Introduction	294
2. Schistosomes as a 'Model' for Anthelmintic MDA	297
2.1 Micro-evolutionary processes	299
2.2 Macro-evolutionary processes	300
3. The Role of Genetics and Genomics in Mathematical Models	301
3.1 Elucidate the basic biology and transmission potential of schistosomes	301
3.1.1 Population structure	301
3.1.2 Transmission rates	304
3.1.3 Spatial heterogeneity	305
3.2 Evaluate transmission cut-off points and estimate when to stop MDA	306
3.3 Elucidate the potential evolution and spread of PZQ resistance	307
3.3.1 Using population genomics to understand the mode of PZQ resistance	311
3.4 Elucidate the potential role of non-human definitive hosts	313
3.5 Elucidate the potential role of hybridization and introgression	314
4. Conclusions	316
Acknowledgments	317
References	317

Abstract

As national governments scale up mass drug administration (MDA) programs aimed to combat neglected tropical diseases (NTDs), novel selection pressures on these parasites increase. To understand how parasite populations are affected by MDA and how to maximize the success of control programmes, it is imperative for epidemiological,

Advances in Parasitology, Volume 87
ISSN 0065-308X
<http://dx.doi.org/10.1016/bs.apar.2014.12.006>

© 2015 Elsevier Ltd.
All rights reserved.

293

Appendix 3

First page from Crellen *et al.* (2016) publication in *Scientific Reports*. Full text available at <http://www.nature.com/articles/srep20954>

OPEN

Whole genome resequencing of the human parasite *Schistosoma mansoni* reveals population history and effects of selection

Received: 08 October 2015

Accepted: 13 January 2016

Published: 16 February 2016

Thomas Crellen^{1,2,4}, Fiona Allan³, Sophia David², Caroline Durrant², Thomas Huckvale², Nancy Holroyd², Aidan M. Emery³, David Rollinson³, David M. Aanensen^{1,2}, Matthew Berriman², Joanne P. Webster^{1,4} & James A. Cotton²

Schistosoma mansoni is a parasitic fluke that infects millions of people in the developing world. This study presents the first application of population genomics to *S. mansoni* based on high-coverage resequencing data from 10 global isolates and an isolate of the closely-related *Schistosoma rodhaini*, which infects rodents. Using population genetic tests, we document genes under directional and balancing selection in *S. mansoni* that may facilitate adaptation to the human host. Coalescence modeling reveals the speciation of *S. mansoni* and *S. rodhaini* as 107.5–147.6KYA, a period which overlaps with the earliest archaeological evidence for fishing in Africa. Our results indicate that *S. mansoni* originated in East Africa and experienced a decline in effective population size 20–90KYA, before dispersing across the continent during the Holocene. In addition, we find strong evidence that *S. mansoni* migrated to the New World with the 16–19th Century Atlantic Slave Trade.

Schistosoma mansoni is a dioecious trematode (flake) and an aetiological agent of the neglected tropical disease schistosomiasis, which infects over 250 million people and causes over 11 thousand deaths annually^{1,2}. The trematode has a wide geographic range; the majority of *S. mansoni* infections are found in sub-Saharan Africa and Madagascar, though transmission foci also exist in the Arabian Peninsula, South America and the Caribbean. Chronic pathology is caused when eggs laid by the adult worms, residing in the mesenteric veins, are swept by the bloodstream into internal organs causing an inflammatory response that may result in fibrosis and calcification of the liver and spleen³. *Schistosoma mansoni* has a complex lifecycle with an intermediate freshwater-snail host (*Biomphalaria* spp.), consequently its epidemiology is closely tied with water contact and prevalence of the disease is greatest in communities that live close to endemic freshwater lakes and rivers⁴.

Since the sequencing of the 380 megabase (Mb) *S. mansoni* reference genome⁵, genomic resources have contributed to comparative⁶ and functional studies⁷, including the discovery of the molecular basis for oxamniquine resistance⁸, but little is known about genome-wide variation between parasites in the field. While large populations of *S. mansoni* have been investigated^{9–11}, all previous studies in this field have been restricted to analysing a small number of markers, typically using the mitochondrial *cox1* gene, nuclear ITS region or <10 microsatellite loci. In this study we present the first population genomic study of *S. mansoni* based on whole genome re-sequencing data for nine isolates collected from field sites across Africa and the New World. African individuals were selected from distinct clades based on previous phylogenetic data¹⁰ to represent the diversity of extant *S. mansoni*. We also include the NMRI (Puerto Rican) laboratory strain that was used to build the reference genome⁵. In addition we have sequenced *Schistosoma rodhaini*, a rodent schistosome species that is firmly established as the closest outgroup to *S. mansoni*^{12,13}.

¹Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, United Kingdom. ²Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, United Kingdom. ³Department of Life Sciences, Natural History Museum, London, SW7 5BD, United Kingdom. ⁴Department of Pathology and Pathogen Biology, Royal Veterinary College, University of London, Hertfordshire, AL9 7TA, United Kingdom. Correspondence and requests for materials should be addressed to T.C. (email: tc13@sanger.ac.uk) or J.A.C. (email: jc17@sanger.ac.uk)

SCIENTIFIC REPORTS | 6:20954 | DOI: 10.1038/srep20954

1

Appendix 4

First page from Crellen, Walker *et al.* (2016) publication in *Clinical Infectious Diseases*. Full text available at <http://cid.oxfordjournals.org/content/63/9/1151.full.pdf>

Reduced Efficacy of Praziquantel Against *Schistosoma mansoni* Is Associated With Multiple Rounds of Mass Drug Administration

Thomas Crellen,^{1,2,3*} Martin Walker,^{1,*} Poppy H. L. Lamberton,^{1,4} Narcis B. Kabatereine,⁵ Edridah M. Tukahebwa,⁵ James A. Cotton,² and Joanne P. Webster^{1,3}

¹Department of Infectious Disease Epidemiology and the London Centre for Neglected Tropical Disease Research, Imperial College London, St Mary's Campus, ²Wellcome Trust Sanger Institute, Hinxton, ³Department of Pathology and Pathogen Biology, Royal Veterinary College, University of London, Hertfordshire, and ⁴Institute of Biodiversity, Animal Health & Comparative Medicine and Wellcome Trust Centre for Molecular Parasitology, University of Glasgow, United Kingdom; and ⁵Division of Vector Borne Diseases, Ministry of Health, Kampala, Uganda

Background. Mass drug administration (MDA) with praziquantel is the cornerstone of schistosomiasis control in sub-Saharan Africa. The effectiveness of this strategy is dependent on the continued high efficacy of praziquantel; however, drug efficacy is rarely monitored using appropriate statistical approaches that can detect early signs of wane.

Methods. We conducted a repeated cross-sectional study, examining children infected with *Schistosoma mansoni* from 6 schools in Uganda that had previously received between 1 and 9 rounds of MDA with praziquantel. We collected up to 12 *S. mansoni* egg counts from 414 children aged 6–12 years before and 25–27 days after treatment with praziquantel. We estimated individual patient egg reduction rates (ERRs) using a statistical model to explore the influence of covariates, including the number of prior MDA rounds.

Results. The average ERR among children within schools that had received 8 or 9 previous rounds of MDA (95% Bayesian credible interval [BCI], 88.23%–93.64%) was statistically significantly lower than the average in schools that had received 5 rounds (95% BCI, 96.13%–99.08%) or 1 round (95% BCI, 95.51%–98.96%) of MDA. We estimate that 5.11%, 4.55%, and 16.42% of children from schools that had received 1, 5, and 8–9 rounds of MDA, respectively, had ERRs below the 90% threshold of optimal praziquantel efficacy set by the World Health Organization.

Conclusions. The reduced efficacy of praziquantel in schools with a higher exposure to MDA may pose a threat to the effectiveness of schistosomiasis control programs. We call for the efficacy of anthelmintic drugs used in MDA to be closely monitored.

Keywords. parasite; schistosomiasis; praziquantel; generalized linear mixed model; anthelmintic efficacy.

Schistosoma mansoni is a parasitic trematode and an etiological agent of the neglected tropical disease schistosomiasis, which is estimated to infect >230 million people worldwide [1]. The parasite has a complex life cycle involving a snail intermediate host, with humans becoming infected through skin contact with infectious water sources. Chronic morbidity of intestinal schistosomiasis is caused by eggs from adult *S. mansoni* that become trapped in tissues. This leads to granuloma formation that can cause fibrosis and portal hypertension [2]. Conservative estimates of morbidity indicate that *S. mansoni* infection accounts for 4.4 million cases of blood in stool and 8.5 million cases of hepatomegaly in sub-Saharan Africa [3].

Interventions to control schistosomiasis have varied over time. Early national control programs, such as the Sudanese

Blue Nile Health Project (started in 1979) and the Egyptian national program (started in 1988), used a combination of chemotherapy, snail control, provision of clean water and sanitation, and infrastructural development [4, 5]. Such multiapproach control programs are now the exception rather than the rule [6]. Due to the low cost, safety, and high efficacy of praziquantel, mass drug administration (MDA) has become the mainstay of national control programs.

MDA began in Uganda in 2003 and 400 000 treatments of praziquantel were distributed that year, with the support of the Schistosomiasis Control Initiative [7–9]. This has since expanded to the treatment of approximately 1.5 million children annually [10], resulting in significant reductions in *S. mansoni* prevalence, intensity, and associated morbidity [11, 12]. Similar successes have been reported in other countries implementing MDA [13]. Spurred by these results, the World Health Organization (WHO) revised its strategic plan from disease control to the interruption of transmission in certain African countries, including Uganda, by 2025 [14]. With the exception of certain islands implementing snail control [15], this ambitious plan remains solely reliant on MDA with praziquantel. Monitoring the continued efficacy of praziquantel is therefore paramount.

Received 3 April 2016; accepted 17 July 2016; published online 28 July 2016.

*T. C. and M. W. contributed equally to this work.

Correspondence: T. Crellen, Department of Infectious Disease Epidemiology, St Mary's Campus, Imperial College London, Norfolk Place, London, W2 1PG, UK (t1c13@sanger.ac.uk).

Clinical Infectious Diseases® 2016;63(9):1151–9

© The Author 2016. Published by Oxford University Press for the Infectious Diseases Society of America. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited. DOI: 10.1093/cid/ciw506