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

Thomas Crellen1,2,4, Fiona Allan3, Sophia David2, Caroline Durrant2, Thomas Huckvale2, Nancy Holroyd2, Aidan M. Emery3, David Rollinson3, David M. Aanensen1,2, Matthew Berriman2, Joanne P. Webster1,4 & 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.
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 genomic methods can identify targets of balancing selection that interact with host immune systems, for instance genes involved in host cell invasion in Plasmodium spp.14–16. Testing for coding sequences under strong positive selection may identify genes involved in longer-term selection in a coevolving host–parasite system17. To shed light on the divergence between S. mansoni and S. rodhaini, we catalogue the extensive genome-wide variation between a global collection of isolates and analyse 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, we look 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 populations18,19. 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 Africa20. 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 region21,22. A subsequent study using a mitochondrial molecular clock suggested that the most recent common ancestor of extant East African S. mansoni occurred 0.3–0.43 MYA and that the divergence between S. mansoni and S. rodhaini was around 2.8 MYA23.

Two main hypotheses have been proposed to explain the divergence between African and South American strains. One suggestion is that African and South American 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 ago24,25. 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 Centuries9. 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 areas21,22.

In contrast to data from a small set of loci, whole genome data offers a more powerful and unbiased approach to reconstruct the history of a species from a coalescent process that gradually builds 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 history and population genetic parameters such as effective population sizes (Ne) and recombination rates26. We used two different coalescent models that both make simplifications in order to approximate the coalescence under recombination from whole-genome sequences26. Pairwise Sequentially Markovian Coalescent (PSMC) uses the pattern of heterozygosity along a single diploid genome to estimate Ne through time27 and Generalised Phylogenetic Coalescent Sampler (G-PhoCS) uses data from 1 kb non-recombining windows across the genome so that the history of population divisions and migration between populations can be modeled28. 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 genomes29, though this is the first study to apply these methods to sequence data from a human-infective parasite.

Results

Genome-wide intra- and inter-species diversity. We sequenced the genomes of 10 Schistosoma mansoni and one S. rodhaini isolate to a median of 46x coverage. Sequence data were mapped against the S. mansoni
reference genome (v5.2); coverage varied between 23x and 58x across 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 of our 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 10× across at least 75% of the coding sequence in the S. rodhaini alignment.

Numbers of SNPs for each sequence, 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 NMRI reference isolate (Supplementary Fig. S1). The total number of SNPs per isolate (including both homozygous and heterozygous sites) ranged from <1 × 10⁶ across the New World and Senegal sequences, to ~1.5 × 10⁶ SNPs in Cameroon and ~2 × 10⁵ SNPs in Uganda. The outgroup species S. rodhaini has ~7 × 10⁶ fixed differences when called against the reference, representing an average of about one variant every 50 basepairs. A notable exception to this geographical trend was the Coastal Kenya sample, which showed a reduced number of SNPs (~8 × 10⁵), more comparable to West African and New World strains than the more divergent Uganda sequences.

Population structure of S. mansoni. We used the identified SNPs across our eleven samples to understand the broad population genetic structure of S. mansoni and their relationship to S. rodhaini. A maximum likelihood phylogeny (Fig. 1) confirmed that S. rodhaini was highly divergent from the ten S. mansoni samples. Rooting the phylogeny with the S. rodhaini sample as an outgroup showed that the deepest split within our S. mansoni samples separated the Uganda isolates (Lake Victoria and Lake Albert, bootstrap support (BS) = 100) from our other samples, followed by the Cameroon samples (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 (Fig. 2). 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.

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, therefore our whole-genome data provides a unique opportunity to identify genes that are potentially host interacting in S. mansoni. We examined positive and negative selection between African S. mansoni and S. rodhaini by comparing dN/dS, using a branch model and balancing selection within S. mansoni using the
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 we set the branch groups as African \(S.\) mansoni and \(S.\) rodhaini using the phylogenetic relationship in Fig. 1. The branch \(d_N/d_S\) values were centered around 1 (interquartile range: 0.44–1.1), examining the tails of the distribution we found 767 coding sequences showing evidence of positive selection (\(d_N/d_S > 2\)) and 2167 genes under strong purifying selection (\(d_N/d_S < 0.1\)). Detailed functional annotation is lacking for most genes in \(S.\) mansoni, nonetheless we explored whether we could 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\(^{30}\), implicating these genes in adaptation to the definitive human host. We 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 we could interpret these results, as 46 out of 52 significant coding sequences were unclassified. Despite this, we 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\(^{32,33}\). The full results are shown in Supplementary Table S1.

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 our results. Furthermore, methods based on modeling the codon substitution process have been shown to be more robust to saturation than alternative methods\(^{34,35}\).

**Hudson-Kreitman-Aguade & McDonald-Kreitman Tests.** The Hudson-Kreitman-Aguade (HKA) test measures the ratio of polymorphisms (\(\pi\)) 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 Fig. 3 and below).

The McDonald-Kreitman skew (MKS) tests for an enrichment of non-synonymous mutations in either fixed differences indicative of directional (positive) selection (MKS < -1) or polymorphisms, indicative of balancing selection (MKS > 1) as measured by a 2 x 2 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

---

**Figure 2. Principal component analyses of isolate SNP data.** (A) including both \(S.\) mansoni and \(S.\) rodhaini samples (Variance represented by PC1: 72.5%, PC2: 9.9%) (B) \(S.\) mansoni samples only, showing more details of the relatedness of isolates within this species (Variance represented by PC1: 36.0%, PC2: 13.7%).
directional selection and 237 genes showed evidence of balancing selection (Supplementary Table S2). 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 (Fig. 3), included in these are heat shock 40 protein (\( hsp40 \)), a homeobox protein, Elongator Complex Protein and an uncharacterised protein (full results shown in Supplementary Table S2).

**Historical demography of *S. mansoni***. Based on evidence from our phylogenetic analysis and PCA we constructed a putative population history for our samples using a number of models (Supplementary Fig. S2). We included African *S. mansoni* and *S. rodhaini* as two populations to estimate the divergence date between species, and effective population size (\( N_e \)) before and after speciation. These posterior values are estimated through coalescent models as functions of two parameters; mutation rate and generation time. Our estimates of these parameters are discussed in the Methods and the Discussion.

Analysis of this model using G-PhoCS 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) 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 = \( 5.53 \times 10^{-2} \)). Ancestral *N_e* before speciation (\( N_e = 4.25 \times 10^3 \)) 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 our models (Supplementary Fig. S2), which contrasts with recent evidence of hybridisation between natural populations of the two species in East Africa\(^{36,37}\), and suggests that these events have not occurred at a high enough frequency over the past 126.5KYA to be detected by our models.

We modeled the split between New World (Guadeloupe) and West African (Senegal and Cameroon) samples (Supplementary Fig. S2). We excluded the NMRI sample from this analysis as it has been kept as a laboratory isolate for hundreds of generations since it was isolated in the early 1940s\(^{38}\) and so violates the assumptions of the G-PhoCS model. The model gives a best estimate of the date of separation 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\(^{39}\), is therefore included within by the 95% confidence interval. Different models 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 (Fig. 2) and G-PhoCS evidence, we model the process with the Ugandan isolates as the basal clade, with
Cameroon and Senegal splitting off (Supplementary Fig. 2). We date the split between Uganda and Cameroon 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 we infer that our \(S.\ mansoni\) and \(S.\ rodhaini\) alleles were present in the same ancestral population (Fig. 4, bootstrapping shown in Supplementary Fig. S3). 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, suggesting that \(S.\ mansoni\) existed as a single population until this point, approximately matching our 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. We also note that PSMC cannot help understand contemporary changes in population size. While our PSMC analysis 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 (Supplementary Fig. S3).

Discussion

We estimate the split between the New World (Guadeloupe) and West African populations (Senegal and Cameroon) to be between 1117–1742 AD. The two proposed possible timings of the Africa –New World split are when the continents split 80–120 million years ago or during the Slave Trade of the 16–19th Centuries. The confidence interval we have obtained spans the historical date of the earliest slaves to the New World. Slaves from French Colonies in West Africa were exported to islands in the French Caribbean, including Guadeloupe.

<table>
<thead>
<tr>
<th>Population</th>
<th>Mean (N_e) (× 10^4)</th>
<th>95% CI (× 10^4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCA (S.\ mansoni)/ (S.\ rodhaini)</td>
<td>42.8</td>
<td>38.7–46.7</td>
</tr>
<tr>
<td>Ancestral (S.\ mansoni)</td>
<td>10.2</td>
<td>9.77–10.7</td>
</tr>
<tr>
<td>Present (S.\ rodhaini)</td>
<td>1.47</td>
<td>1.26–170</td>
</tr>
<tr>
<td>Present East Africa (S.\ mansoni)</td>
<td>6.50</td>
<td>3.67–9.35</td>
</tr>
<tr>
<td>Present West Africa (S.\ mansoni)</td>
<td>0.700</td>
<td>0.330–1.09</td>
</tr>
<tr>
<td>Present Guadeloupe (S.\ mansoni)</td>
<td>0.307</td>
<td>0.147–0.473</td>
</tr>
</tbody>
</table>

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 4. 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 show in Supplementary Fig. S3.
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. Our result shows the data from whole genome sequences to be consistent with both historical dating and the results found from previous studies on \( S. \text{mansoni} \) using fewer 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, though the role of humans in transmitting diseases between continents is contested. For instance, \( M. \text{tuberculosis} \), 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 \( M. \text{tuberculosis} \) from Old to New World. Diseases such as syphilis, caused by the spirochete \( T. \text{pallidum} \), are assumed to have been transmitted to the New World with the Conquistadors of the 16th Century and this is supported by archaeological findings, though molecular data is lacking. The protozoan parasite \( L. \text{chagasi} \), 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 over 500 years ago, a migration event which caused a genetic bottleneck that has been observed in microsatellite loci. The evidence presented here builds on the 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 \( S. \text{mansoni} \) as a distinct species occurred around 126.5KYA. This estimate is an order of magnitude earlier than previous estimates that placed the split at 2.8MYA and 1.2–10MYA respectively. We propose that this most likely occurred in East Africa as the Ugandan isolates are the earliest diverging \( S. \text{mansoni} \) samples on the maximum likelihood tree. This supports the conclusions drawn from previous mtDNA barcoding studies. We hypothesise that the adoption of fishing by human communities may have triggered the speciation of \( S. \text{mansoni} \) and \( S. \text{rodhaini} \) as this led to hunter-gatherer groups residing for longer periods by water bodies, leading to the establishment of \( S. \text{mansoni} \) 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). The earliest evidence is from Mossel Bay, South Africa where stone artefacts and substantial shellfish deposits are dated to 152–176KYA. Similarly, at the Red Sea Basin in Eritrea deposits of stone tools and shells from marine organisms have been dated to 118–132KYA. 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. Our posterior distribution for the speciation of the last common ancestor of \( S. \text{mansoni} \) and \( S. \text{rodhaini} \) overlaps with these archaeological dates (Fig. 5), 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-Africans51. The East African climate was generally arid from 60KYA until the beginning of the Holocene52 and fluctuations in temperature and precipitation throughout this period impacted on the size and depth of East African Lakes, where we presume 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 present53. As PSMC is unable to specify exactly when in the 20–100KYA window the bottleneck in *S. mansoni* occurred, we are not able 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 our phylogenetic tree with coalescence modeling we 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. We 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 Regions54. 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 our models is summarised in Fig. 6.

Both coalescence models presented here use two parameters, generation time and mutation rate. We have made estimates of both these values as 0.2 years (see Methods) and $8.1 \times 10^{-9}$ per basepair per generation (Supplementary Fig. S4) respectively55–57. 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 infection58. Given the variability of such factors spatially and temporally, our estimate of the generation time is a simplification. Our estimate of the mutation rate, based on the relationship between genome size and mutation rate, may be subject to alteration on the basis of experimental evidence59. Therefore if revisions are made to either the mutation rate or generation time in the future, both dates of divergence and *N_e* estimates provided here will need revision. Despite these limitations, we find it encouraging that our estimate of the timing of migration to the New World overlaps with historical dating, suggesting that our estimates of the parameters are likely to be robust.

Our sample from coastal Kenya 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 mitochondria9. 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 Mombasa60. Therefore 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.

Through population genetic tests where we specify S. rodhaini as the out-group, we have highlighted genes that show a deviation 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 $d_s/d_o$, we found 767 genes showing evidence of positive selection ($d_s/d_o > 2$); 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 we tested for significance under a likelihood ratio framework, we found VAL21 and a cercarial elastase gene 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 host62. Cercarial elastase cleaves insoluble elastin, a major component of the dermis, and is the only protease required for mammalian host invasion63. While cercarial elastase genes are broadly conserved within the genus Schistosoma, our finding of a signal of positive selection in Smp_119130 may hold the key for human-specific invasion.

Gene with high values in the MKS, suggestive of balancing selection, show enrichment for pfam domains found in immunoglobulin 1-set proteins, a domain linked to cell–cell interactions that could facilitate adaptation to the host. Dynemin-domain containing proteins 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 selection64. As the isolates are not from first passage (see Methods), this represents a potential limitation on our analysis for selection if passages have 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 tapeworms65. While there may be multiple explanations for the pattern of variation observed on any particular gene, the loci we identify 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 our tests for selection can be difficult. Efforts are therefore underway to develop high-throughput functional genomic approaches to flatworm biology63 and the genes that we have identified as outliers in the selection tests from this analysis will be fed into an ongoing RNAi screen to examine function.

Methods and Materials

Samples. Samples (Table 1), with the exception of NMRI, came from the schistosomiasis collection (SCAN) at the Natural History Museum (NHM), London64. 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 generations64 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 Analyzer 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/). ENA Accession numbers for each sample are shown in Table 1.

Variant Calling Pipeline. Reads were mapped against v5.2 of the S. mansoni reference genome65 using smallt (http://www.sanger.ac.uk/resources/software/smallt/) 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 repeatedly mapped reads are excluded by our 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 indels65. Single nucleotide polymorphisms (SNPs) were called using three algorithms; HaplotypeCaller and UnifiedGenotyper (implemented in GATK) and samtools version 0.1.19 mpileup66. 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.
The 11727 coding sequences in the S. mansoni genome were filtered by removing alternative trans- 
cripts (1005 sequences removed), genes with low coverage in the S. rodhaini alignment (<25% of bases covered 
at 10 × 719 sequences removed), genes with excessive coverage (>175 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 dN/dS for 
each protein-coding gene were calculated by running codeml from the PAML package67 in branch mode, where 
the phylogenetic tree for African S. mansoni and S. rodhaini was specified and partitioned into two clades, sepa-
rating the branch leading to the S. rodhaini sequence (the ‘foreground’ branch) from all other branches between 
different S. mansoni samples. We ran two scenarios for each coding sequence, in which the value for dN/dS 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. We calculate the test statistic D, where D = −2(lnLfix 
−lnLunfixed). 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 = paramsunfixed -paramsfixed, for all cod-
ing sequences in our dataset df = 1. We set the significance threshold for the Chi-Square distribution at p < 0.05. 
Statistical overrepresentation analysis was conducted using PantherDB (version 10.0)68.

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 (ie divers-
ity 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 
software69.

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 
software70. The MK skew, or neutrality index, is defined as log2 ((Npol/Spol)/ (Nfix/Sfix)), where Ns 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 PantherDB68 and enrichment of 
pfam domains was performed using a one sided Fisher’s exact test in R (version 3.2.2).

Consensus fasta files. The phylogenetics software package RaxML and coalescence models PSMC and 
G-PhoCS require fasta sequences, so in the absence of phase information for our samples we calculated a consen-
sus sequence of the diploid genotypes. These were 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 chromosomes30: this left the scaffolds aligned to the 7 autosomal chromosomes. To ensure loci included 
sequences, scaffolds were unlinked and sufficiently small that little recombination occurs within loci, we examined 
the decay in linkage disequilibrium (LD) with distance between pairs of SNPs in our data using SNP data in vcf files, 
which were phased using SHAPEIT2 (v2.790)40, and analysed using VCFtools (v0.1.11) –hap-r2 and found that 
LD decays to background levels for distances great than around 400kb (Supplementary Fig. S5). To be conserv-
ative, we thus divided our consensus calls into 1kb blocks separated by 500kb. These same input regions were 
concatenated and used as input for RAxML.

Phylogenetic Tree. We constructed a Maximum likelihood (ML) tree using RAxML version 8.0.2471 under 
a GTR model of nucleotide substitution with a gamma distribution of rate variation across sites72. 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 was visualised in FigTree 
(v1.4.0)72.

Principal Component Analysis. We built a matrix where every variable site in the genome is classified as 
the relationship between log-genome size and log-mutation rate73. The best fitting model for mutation 
rate μ in bp was mu = 2.585324 × 10−10 × g0.384 for genome size g in Mb, giving an estimate of 8.1 × 10−9 per bp 
per generation (Fig. S4). Generation time estimated by adding the time taken for the parasite to start producing 
eggs after infection with human host (~42 days)74 with the latency period in snails (~28 days) + time for tran-
section 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 (Ne) in numbers of individuals, we used the relationship Ne = θ(4πμ) where θ is 
the uncalibrated population size and μ is the mutation rate. To obtain the divergence time in years (D), we used 
the relationship, D = (0.2τ)/μ 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 species28. We ran a 
number of different G-PhoCS analyses to test how sensitive estimates of these parameters are to different assump-
tions about population structure and migration (Fig. S2). 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. 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 “262 + 4 + 6”, with the distribution of TMRCAs 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 (Supplementary Fig. S3).

**References**

We thank Jane Lomax and Adam Reid for making recommendations on GO term analysis, Bonnie Webster for selecting samples for sequencing from the SCAN collection, André Thèron for providing isolates from Université de Perpignan and Thomas Otto for assistance in implementing the tests for selection. TC is funded by an MRC studentship. SD, CD, TH, NH, DMA, MB and JAC are supported by the Wellcome Trust via core support of the Wellcome Trust Sanger Institute (grant 098051).
Author Contributions

Additional Information
Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Crellen, T. et al. Whole genome resequencing of the human parasite Schistosoma mansoni reveals population history and effects of selection. Sci. Rep. 6, 20954; doi: 10.1038/srep20954 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/