Ancient hybridization and introgression of an invadolysin gene in schistosome parasites

Roy N. Platt II[1], Marina McDew-White[1], Winka Le Clec'h[1], Frederic D. Chevalier[1], Fiona Allan[2,3], Aidan M. Emery[2, 3], Amadou Garba[4], Amina A. Hamidou[4], Shaali M. Ame[5], Joanne P. Webster[6, 3], David Rollinson[2, 3], Bonnie L. Webster[2, 3], Timothy J. C. Anderson[1*].

[1] Department of Genetics, Texas Biomedical Research Institute, PO Box 760549, San Antonio, TX 78258

[2] The Natural History Museum, Department of Life Sciences, Cromwell Road, London SW7 5BD, United Kingdom

[3] London Centre for Neglected Tropical Disease Research (LCNTDR), Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, United Kingdom

[4] Réseau International Schistosomoses, Environnement, Aménagement et Lutte (RISEAL-Niger), 333, Avenue des Zarmakoye, B.P. 13724, Niamey, Niger

[5] Public Health Laboratory - Ivo de Carneri, Pemba, United Republic of Tanzania

[6] Centre for Emerging, Endemic and Exotic Diseases, Department of Pathobiology and Population Sciences, Royal Veterinary College, University of London, AL9 7TA, UK United Kingdom

*Correspondence to: tanderso@txbiomed.org.

The parasitic blood fluke *Schistosoma haematobium* causes urogenital schistosomiasis in humans and is a major cause of morbidity and mortality across sub-Saharan Africa. *S. haematobium* hybridizes with livestock schistosomes, including S. *bovis*, however the frequency, direction, age and genomic consequences of hybridization are unknown. We sequenced 96 *S. haematobium* exomes from Niger and the Zanzibar archipelago. and found evidence of an ancient, introgression event between Nigerien *S. haematobium* and *S. bovis* occurring 108-613 generations ago. Between 3.3-8.2% of Nigerien *S. haematobium* genomes are derived from *S. bovis* alleles, some of which show signatures of directional selection; the strongest signal spans a single gene in the invadolysin gene family, an M8 metalloprotease associated with parasitic life-history traits.

Introgressive hybridization occurs when hybrid offspring repeatedly backcross with one or both parental types, acting as a conduit for genetic exchange between species. Once present in the "new" genetic background, introgressed loci are broken up through recombination and exposed to selection. Alleles that are deleterious in the new genetic background are purged while advantageous alleles may increase in frequency if they escape loss by drift. Examples of introgression are increasingly common[4]. Hybridization among parasite species is a concern because genes encoding biomedically important traits such as virulence, host specificity or drug resistance may be transferred between species[5]. Schistosomes are dioecious trematodes that cause the chronic and debilitating disease schistosomiasis in humans and animals. Inter-specific hybridization, and the production of viable hybrid offspring, have been demonstrated experimentally between members of the *S. haematobium* species group[6]. Adult worms live in the blood vessels of their human hosts and cannot be easily sampled directly. However, there is genetic evidence --discordance between nuclear (ITS rDNA) and mitochondrial (cox1) DNA-- for natural hybridization of *S. haematobium* with closely related sister species in Africa[7-11] from genotyping larval parasites. One such hybridization of major concern is between *S. haematobium* and the livestock schistosome, *S. bovis* commonly observed in West Africa.

To investigate this natural hybridization at the genomic level, we collected *S. haematobium* eggs, from human urine samples, and hatched miracidia. We sequenced exomes from 96 miracidia (1 per patient) from Niger (n=48) and Zanzibar (n=48) using whole genome amplification and exome capture (Data S1; Fig S1; [12)]. Initial attempts to genotype mitochondrial loci contained within the exome probe set failed in 31 of the Nigerien S. *haematobium* samples. We therefore used *de novo* assembly of the mitochondrial genome and *cox1* Sanger sequencing to genotype the mitochondrial DNA and found that these samples contained a *S. bovis* mitochondrial haplotype that was 15-18.05% divergent from that of *S. haematobium* and therefore poorly captured by our exome capture baits (fig S2; [13)]. In all, 65% (*n*=31) of the Nigerien *S. haematobium* miracidia examined had a *S. bovis* mitotype, confirming *S. bovis* mitochondrial introgression into *S. haematobium* as previously identified in other West African *S. haematobium* populations[9,11]. In contrast, the *S. bovis* mitotype was not present in the Zanzibari *S. haematobium* population.

The presence of *S. bovis* mitochondria in Nigerien *S. haematobium* could result from contemporary hybridization or past introgression events. To differentiate between these two

scenarios, we examined 370,770 autosomal, exonic (single nucleotide polymorphisms) SNPs[12,14] recovered by the exome capture probes. In addition to our 96 *S. haematobium* samples we used sequence data from six closely-related schistosome species in the *S. haematobium* species group (*S. bovis*, *S. curassoni*, *S. intercalatum*, *S. guineensis*, *S. mattheei*, and *S. margrebowiei*) generated in previous studies[15,16]. The genomes of *S. haematobium* and *S. mansoni* are ≥89.4% syntenic[15]. To take advantage of the more contiguous, chromosomal length assembly and gene annotations available for *S. mansoni*[17,18], we aligned the *S. haematobium* and *S. mansoni* assemblies to convert the coordinates from their position on *S. haematobium* contigs to the corresponding position on *S. mansoni* chromosomes. We verified synteny by comparing linkage disequilibrium (LD) between the two sets of SNP positions (fig S3).

A principle component analysis (PCA) comparing all *Schistosoma* samples clearly differentiated *S. haematobium* from all other species along the first two PCs which accounted for 48.9% of genotypic variation observed (Fig 1A). The lack of intermediate genotypes between *S. haematobium* and other schistosome species suggests we did not have early generation hybrids within our samples. We used ADMIXTURE[19] to assign ancestry proportions to the *S. haematobium*, *bovis*, and *curassoni* samples (table S1). *S. curassoni* was included due to its close phylogenetic affiliation with *S. bovis*. Three distinct ancestry components were identified corresponding to *S. bovis*/*curassoni*, Nigerien, and Zanzibari *S. haematobium*. The *S. bovis*/*curassoni* ancestry component was present in 16 of the 48 Nigerien *S. haematobium* miracidia indicating potential low levels of admixture ($0.1 < Q > 2.7$; Fig 1B), while no admixture was identified in the Zanzibari *S. haematobium*. The three-population test (f3;[20)] indicated admixture in the Nigerien *S. haematobium* population and identified *S. bovis* as the most likely source (Fig 1C). Finally, we confirmed that *S. bovis* alleles in the Nigerien *S. haematobium* samples were the result of introgression as opposed to incomplete lineage sorting with Patterson's *D* (D = 0.144, standard error = 0.04; [21)]. Combined, these results indicate past introgression event(s) between Nigerien *S. haematobium* and *S. bovis*. We used PCAdmix[22] to identify introgressed alleles in the Nigerien *S. haematobium*. Ancestry was assigned to haplotype blocks by comparing them to a reference panel of *S. bovis* and Zanzibari *S. haematobium*. Within each Nigerien *S. haematobium* miracidium 3.3-8.1% (mean = 5.2%) of alleles were identified as *S. bovis* haplotypes. In contrast, we observed close to zero (0.004 - 0.2%) *S. bovis* ancestry in a subsample of five Zanzibari miracidia.

Recombination breaks down introgressed segments over time. As a result, the length of introgressed haplotype blocks can be used to date time since admixture in number of recombination events/generations[23]. Using this method, admixture occurred ~240.6 generations ago (min = 107.8 max = 612.5; fig S4). Estimates of *Schistosoma* generation time vary greatly: while adult worms can live up to 37 years[24], the minimum time for eggs to mature into adult worms is ~120 days. Assuming a ~1 year generation time[25] the hybridization event leading to introgression occurred ~240.6 years ago (min = 107.8 max = 612.5). Our time since introgression is likely conservative since we have limited ability to identify small haplotype blocks which would underestimate our time since introgression. Despite these limitations the genomic data demonstrate that the introgression event occurred more than a hundred generations ago but subsequent to the divergence of the Nigerien and Zanzibari *S. haematobium* populations. While all 48 Nigerien miracidia contain introgressed *S. bovis* DNA, we observed no F1 or early generation hybrids.

Since the admixture event some *S. bovis* alleles are approaching or have become fixed in the Nigerien *S. haematobium* population (Fig 2) through positive selection. We used two complementary methods to examine directional selection in the Zanzibari and Nigerien *S. haematobium* samples. BayeScan[26] was used to quantify selection (α) as a function of allele frequency differences at individual SNPs while larger haplotype blocks under directional selection were identified using cross-population extended haplotype homozygosity (xpEHH; fig S5; [27)]. BayeScan and xpEHH identified 2 and 17 regions under directional selection between the Zanzibari and Nigerien *S. haematobium* populations. The strongest signals of selection from both analyses ($\alpha > 1.75$; xpEHH ≤ -3) and highest frequency (100%) of introgressed *S. bovis* haplotype blocks in the Nigerien *S. haematobium* population occurred at a 23Kb locus on chromosome 4 (Chr4:20,023,951-20,047,325; Fig 2). To confirm that this locus was the product of an introgression event rather than the result of selection on standing variation, we calculated the time to the common ancestor of the 23 Kb locus in the Nigerien haplotypes relative to the surrounding genomic regions using startmrca[28]. Using mutation and recombination rates[29] reported for *S. mansoni*, we estimated the time since divergence of the 23 Kb target region in the Nigerien haplotypes to be 476 (437.9-616.6 95% credible interval) generations ago, well within the range of our previous estimates of introgression from the haplotype block length analysis.

Further examination of this 23 KB region revealed that it spanned a single gene, Smp_127030, alternatively referred to as invadolysin. Within the invadolysin gene, we found 23 SNPs coding for 14 nonsynonymous changes (excluding singletons) two of which show fixed differences between Zanzibar and Niger populations (fig S6). Invadolysin is a member of an gene family within the M8 metalloproteases originally identified in *Leishmania*[30]. In *Leishmania*, these are surface proteins associated with degradation of collagen and fibrenonectin in the extracellular matrix and larval penetration of mammalian hosts[31]. In schistosomes M8 metalloproteases have undergone a rapid gene family expansion[16,32,33] (fig S7) and are among the most abundant transcripts and secreted proteins[33] in larval stages. RNAi knockdowns of the invadolysin paralog in *S. mansoni* miracidia led to reduced larval penetration and establishment in the snail intermediate host, reducing cercariae production[34].

The six amino acid changes we were able to place on a protein structure of Smp_127030, modeled from the *Drosophila* invadolysin (Protein DB: d1lmla), were in peripheral positions on the protein (fig S8). Members of the invadolysin family are expressed in stage specific manners, similar to globins in vertebrates[35]. In *S. mansoni* the Smp_127030 paralog is expressed during stages associated with the mammalian host and most highly expressed in adult worms (fig S9). Smp_127030 is expressed in adult worms in *S. haematobium*, but expression data is not available for other life stages (except eggs). It is unclear how the *S. bovis* invadolysin (Smp_127030) allele benefits *S. haematobium*. We speculate that the introgressed invadolysin is involved in tissue penetration or immune evasion within the mammalian host.

Hybridization between species is a powerful source of evolutionary novelty. Introgression among parasites is a concern[5], because genes conferring biomedically important traits may be transferred across species boundaries; this has clearly taken place in *S. haematobium*. While we see no evidence of early generation hybrids, we see a signature of introgression from *S. bovis* in all Nigerien miracidia examined. We demonstrate that an introgressed M8 metalloprotease (invadolysin; Smp_127030) expressed in adult worms has spread to high frequencies in *S. haematobium* populations in Niger, and that parasites bearing introgressed *S. bovis*-derived alleles have a strong selective advantage over those carrying native *S. haematobium* alleles.

Materials and Methods

Ethics statement: For the Niger sample collection, ethical clearance was obtained from the Niger National Ethical Committee, in combination with the St Mary's Hospital Local Ethics Research Committee (part of the Imperial College London Research Ethics Committee (ICREC; (EC NO: 03.36. R&D No: 03/SB/033E)) in London, United Kingdom, in combination with the ongoing Schistosomiasis Control Initiative (SCI) and Schistosomiasis Consortium for Operational Resaerch (SCORE) activities. For the Zanzibar sample collection, ethical approval was obtained from the Zanzibar Medical Research and Ethics Committee (ZAMREC, reference no. ZAMREC 0003/Sept/011) in Zanzibar, United Republic of Tanzania, the "Ethikkomission beider Basel" (EKBB, reference no. 236/11) in Basel, Switzerland, the Institutional Review Board of the University of Georgia (project no. 2012-10138-0), and registered at the International Standard Randomised Controlled Trial (Register Number ISRCTN48837681). Within both Niger and Zanzibar, all aspects of sample collections were carried out in the framework of the disease control activities implemented and approved by the local Ministry of Health (MoH) and adopted by regional and local administrative and health authorities.

The study participants were informed about the study objectives and procedures. Written consent was obtained from parents prior to sample collection from children. Participation was voluntary and children could withdraw or be withdrawn from the study at any time without obligation. All children were offered PZQ (40 mg/kg single oral dose) treatment in the frame of the following school-based or community-wide treatment carried out by the MoH.

Sample collection: This study used archived miracidia samples from Niger and Zanzibar (Tanzania) fixed on Whatman FTA cards archived within the Schistosome Collection at the Natural History Museum (SCAN)[36]. From Zanzibar, encompassing both the Unguja and Pemba islands, the *S. haematobium* miracidia were collected as part of the SCORE population genetics studies in 2011 from 26 locations spaced up to 160.9 km apart. From Niger the *S. haematobium* miracidia were collected in 2013 from school-aged children from 10 locations located up to 125 km apart. These samples were also collected as part of SCORE population genetic studies within a gaining and sustaining control of schistosomiasis project in Niger and also as part of monitoring and evaluation activities carried out by the Schistosomiasis Control Initiative. To capture maximum diversity we used a single miracidium from 96 individuals (n_{Zanzibar}=48, n_{Niger} = 48). *S. haematobium* eggs were harvested from each infected urine sample by sedimentation or filtration,

put into fresh water and then exposed to light to facilitate miracidial hatching. Miracidia were captured individually, under a binocular microscope, in 3 µL of water and spotted onto indicating Whatman FTA Classic Indicating cards (GE Healthcare Life Sciences, UK) using a 20 µL micropipette, dried and archived in SCAN[36].

Library prep and sequencing: We used whole genome amplification of single miracidia dried on FTA cards followed by exome capture and sequencing (Illumina 2500) following methods described in [12]. In addition to exome capture, whole-genome sequence data was generated for twelve samples, six from each population (Niger and Zanzibar). Libraries were multiplexed and paired end 150 bp reads were sequenced on a single Illumina NextSeq flowcell. In addition to the whole genome and exome sequence data generated above, we gathered available genome sequence data from the NCBI Short Read Archives for six other species of *Schistosoma* in the *haematobium* group, including *S. bovis* (ERR119622, ERR103048, ERR539853), *S. curassoni* (ERR310937, ERR119623), *S. haematobium* (ERR084970, ERR037800, SRR433865), *S. intercalatum* (ERR539854, ERR539856, ERR119613), *S. guineensis* (ERR119612, ERR539850, ERR539852), *S. mattheei* (ERR539851, ERR539855, ERR539857, ERR103051), and *S. margrebowiei* (ERR310940) in 20 separate libraries[15,16].

Computational environment: All analyses were conducted on a high-performance computing cluster within a Singularity container or Conda environment. Environmental recipe files, custom programs, and shell scripts are at https://github.com/nealplatt/sH_hybridization (v1.0; doi: 10.5281/zenodo.2536390).

Variant discovery, filtration, and phasing: Sequence reads were trimmed with Trimmomatic v0.36 [37] so that the leading and trailing base calls had phred-scaled quality scores greater than 10 and the phred score was greater than 15 over a 4 nt sliding window. After trimming, paired and singleton reads were mapped to the reference *S. haematobium* genome (SchHae_1.0, GCA_000699445.1, lab strain, originally from Egypt). BWA v0.7.17-r1188[38] and merged into a single BAM file using SAMtools v1.7[39]. GATK v4.0.1.1[40] was used to add read group information and mark duplicate reads from each library. Where possible, complete read group information was added based on information contained within the FASTQ header. In some cases (primarily the public data) pseudo read groups were created to differentiate each library.

GATK's Haplotypecaller[40] was used for variant discovery. Variant discovery was restricted to target regions of the *S. haematobium* assembly using the –L option. Target regions were identified by combining all adjacent, exome probe locations within 500 bases of one another into larger intervals with BEDtools v2.27.1[41]. Each interval was genotyped using GATK's GenotypeGVCFs and combined into a single VCF for filtering using GATK's MergeVcfs. Low quality SNP genotypes were filtered using GATK's VariantFiltration with the following filters: variant quality normalized by depth (QD < 2.0), strand bias (FS > 60.0), mapping quality (MQ < 40.0), mapping quality rank sum (MQRankSum < -12.5), and read position rank sum (ReadPosRankSum < -8.0). Sites with high rates of missing data (>20%), multi-allelic sites, and individuals with low call rates (>15% missing data) were removed using VCFtools v 0.1.15[42]. All indels were excluded from downstream analyses.

The published *S. haematobium* and *S. mansoni* assemblies vary greatly in quality, despite a high degree of synteny between the two genomes[15]. We aligned the two genomes using progressiveCactus v0.0[43] using default parameters to leverage the contiguity of the *S. mansoni* assembly. The HAL alignment file was used to lift SNP coordinates between the assemblies using progressiveCactus' halLiftover module. Multi-position SNPs, those that align between assemblies in something other than a 1:1 relationship, were removed from downstream analyses. We used linkage disequilibrium (LD) decay curves to examine biases introduced during the liftover by comparing LD from SNPs associated with each assembly. SNPs mapping to the *S. mansoni* autosomes (chr1-7) were extracted using VCFtools v0.1.15. The square of the correlation coefficient (r^2) was calculated for all SNPs within 1.5 Mb of each other for each dataset using PLINK v1.90b4[44]. The 1.5 Mb cutoff was used since it represented the largest scaffold in the *S. haematobium* assembly. After confirming concordance between the original and *S. mansoni*-lifted datasets, we used Beagle v4.1[45] to impute missing SNPs and phase haplotypes for each autosomal chromosome. Imputations and phasing occurred in sliding windows of 300 SNPs and a step size of 30 with 250 iterations per window.

<u>*Mitotype assignment*</u> – Unique filters were used to manage the mitochondrial SNPs. First, the mitochondrial contig was identified from the *S. haematobium* assembly using the *haematobium* vs. *mansoni* whole genome alignment and confirmed using NCBI's BLAST server. Due to low genotyping rates within the Nigerien samples, mitochondrial SNPs were filtered so that the

genotyping rate per site was reduced from 85% to 25%. Putative mitochondrial haplotypes (mitotypes) were generated by removing any heterozygous sites, if present. Given previously described rates of mitochondrial introgression and divergence between *S. bovis, S. curassoni* and *S. haematobium,* we mapped filtered reads directly to the *S. haematobium* mitochondrial contig (AMPZ01026399.1) and a previously generated *S. curassoni* mitochondria sequence (AP017708.1). Full-length *S. bovis* mitochondrial genomes are not currently available through NCBI Genbank (last accessed 23 April 2018). *S. bovis*/*curassoni* or *haematobium* mitotypes were classified using the ratio of reads mapping to each mitochondrial reference and confirmed by Sanger sequencing the mitochondrial, cytochrome C oxidase 1 (*cox1*) gene. *cox1* was amplified in a reaction containing 1x reaction buffer, 0.8 mM dNTPs, 1.5mM MgCl2, 1 µM forward primer (Cox1_schist_5'; CTT TRG ATC ATA AGC G; [46]), and 1 µM reverse primer (Cox1_schist_3'; TAA TGC ATM GGA AAA AAA CA; [46]) and under the following reaction conditions: 2 minute (min), 95˚C initial denaturation, 35 cycles of a 30 second (sec), 95˚C denaturation, 30 sec, 52˚C annealing phase, and a 1 min 72˚C extension, followed by a final 7 min, 72˚C extension cycle. Amplification and fragment size were confirmed on a 1.5% agarose gel. Bidirectional sequences were generated for each fragment with the Cox1_schist_5' and Cox1_schist_3' primers by the Eurofins (Eurofins-MWG) sequencing service. All *cox1* sequences were compared to available sequences on GenBank for species identification.

Population structure, admixture, and ancestry assignment - Summary statistics were calculated using all filtered autosomal SNPs with minor allele frequency greater than 5%. F_{IS} was calculated for each sample using VCFtools and the "--het" option. The scikit-allel v1.1.10[47] Python library was used to calculate F_{ST} between the Zanzibari and Nigerien populations of *S. haematobium*. Genome-wide values for F_{ST} were averaged from blocks of 100 variants and local calculations were generated from sliding windows of 250 kb and 50 kb steps. Relationships between samples were visualized in genotypic space with a PCA of unlinked SNPs. Linked sites within sliding windows of 25 SNPs and a pairwise R^2 value greater than 0.2 were filtered using PLINK v1.90b4[44]. Since comparison between species can overwhelm population level clustering, separate PCAs were generated from previously published *Schistosoma* species[15,16], and the exome data from the *S. haematobium* miracidia (Niger and Zanzibar) presented herein. Nucleotide diversity (π) was calculated per site and in sliding windows of 50 Kb with 25 Kb steps using VCFtools.

Levels of admixture were calculated using ADMIXTURE v1.3.0[19]. The number of populations (*K*) was explored from 1 to 20 and cross-validation between 1,000 replicates was used to identify the most robust measure of *K*. The 3-population test f_3 as implemented in the scikit-allel was used to identify admixture in the Nigerien and Zanzibar *S. haematobium* populations[20]. f_3 standard errors and *Z*-scores were calculated from block-jackknife replicates of 100 SNPs. We used the scikit-allel package to calculate *D* statistics and compare possible introgression from *S. bovis* or S. *curassoni* into *S. haematobium*. Standard error and *Z*-scores were calculated by block-jackknife replicates in windows of 100 SNPs. In addition to genome-wide averaged, *D* was calculated in local, sliding windows of 50 SNPs with 25 SNP steps. Finally, PCAdmix v1.0[22] was used for SNP ancestry assignment. For individuals in the admixed Nigerien population, ancestry was assigned in sliding windows across each chromosome to one of two parental populations defined by "pure" *S. haematobium* (represented here by samples from Zanzibar) and *S. bovis*. For validation, we included five, randomly selected, Zanzibari *S. haematobium* samples with the assumption that the entire genome should be assigned Zanzibari ancestry. Ancestry was assigned to windows of 30 SNPs at a time, with a minimum assignment threshold of 99%.

Selection– We quantified selection across the *S. haematobium* exome using two independent methods. BayeScan v2.1[26] proposes two demographic models to explain allele frequency differences between at each locus. The first model assumes allele frequency differences are due to drift while the other model adds an additional parameter (α) to account for selection. A Markov-chain Monte Carlo (MCMC) was used to generate a posterior distribution of the alpha parameter such that significant deviations from zero indicate directional selection. We examined selection by comparing the Zanzibar and Niger *S. haematobium* populations using 1,000 separate BayeScan runs (chains) to guarantee convergence on posterior distributions alpha. For each chain we used 50 pilot runs of 10,000 generations to generate starting priors and burnin time of 50,000 generations. Prior odds for the neutral models ($\alpha = 0$) were set at 10:1. To minimize autocorrelation, samples were thinned by 20 and 50,000 samples were taken. Convergence between chains was examined in R using the *coda* package v0.19-1. A Gelman-Rubin convergence diagnostic score less than 1.1 was defined as an acceptable level of convergence *a priori*. In addition to BayeScan, we identified regions under selection in both the *S. haematobium* populations using cross population extended haplotype homozygosity (xpEHH; [27)] score as implemented in the *rehh* v2.0.2 R package[48]. When possible, alleles were polarized against an

outgroup (*S. margrebowiei*) to define the ancestral state. If sites were heterozygous or missing in *S. margrebowiei* they were excluded. Regions under selection were defined using a $-\log_{10}(p\text{-value}) = 3$ an an $|xpEHH| > 3$. Adjacent loci ≤12.5 kb apart were combined into a single locus.

<u>*Dating genome wide introgression*</u> – *S. bovis* introgression tracts were identified using Python scripts available at https://github.com/nealplatt/sH_hybridization (last accessed 21 January 2019). Briefly, derived (autapomorphic) *S. bovis* and Zanzibari *S. haematobium* alleles were identified and used to annotate Nigerien *S. haematobium* alleles. Introgression blocks were identified by finding the largest stretch of *S. bovis* or pleisiomorphic alleles that were unbroken by a Zanzibari *S. haematobium* allele. With these tracts we estimated the number of generations since admixture[23]:

$$G = \frac{1}{L * P}$$

Where G is generations, L is the average length of introgression tracts in Morgans and P is the proportion of the genome from the major parent.

We dated time since divergence of the Chr4:20,023,951-20,047,325 locus in the Nigerien *S. haematobium* populations using startmrca[28]. We used a uniform recombination rate of 3.4×10^{-8} and mutation rate of 8.1×10^{-9}. We centered the analysis on Chr4:20,033,013 and included 1Mb of upstream and downstream sequence. MCMC chains were run for 50,000 generations and limiting proposals to 20 standard deviations. Ten independent chains were run using the parameters described above. Estimates of divergence time were taken by discarding the first 40,000 generations of each run then thinning the remainder to 1 sample per 10 generations.

<u>*M8 peptidase and Invadolysin gene family evolution*</u> – Invadolysin (Smp_127030) is a member of the M8 peptidase family of proteins (Pfam ID: PF01457) from the Pfam database[49] (v32.0; accessed 28 October 2018). We downloaded amino acid sequences for all platyhelminths and several outgroup taxa including *Homo sapiens*, *Mus musculus*, *Monodelphis Domestica*, *Gallus gallus*, *Anolis carolinesis*, *Danio rerio*, *Drosophila melanogaster*, *Aedes* aegypti, and *Caenorhabditis elegans*. Sequences missing the HExxH active site[50] were removed from downstream analysis. Amino acid sequences were aligned with MUSCLE v3.8.1551[51], and sites with less than 75% coverage were removed. We used RAxML v 8.2.10[52] to complete 100 searches

for the optimal tree using the PROTGAMMAWAG substitution model. Gene duplication evens were inferred using Mega7[53] and the gene duplication wizard. Duplications were categorized based on their presence in lineages leading to *Schistosoma* paralogs.

Invadolysin (Smp_127030) structure – We used Phyre2[54] (http://www.sbg.bio.ic.ac.uk/~phyre2; accessed 15 November 2018) in "intensive mode" to model the Nigerien *S. haematobium* invadolysin (Smp_127030) protein structure. We used the whole genome resequencing data to identify SNPs at the invadolysin (Smp_127030) locus since the exome probes only targeted 76.1% of the coding region (1,722 of 2,262 bp). Pockets in the protein structure were identified with fpocket2[55] as implemented in PHYRE2 web portal. After modeling the invadolysin (Smp_127030) protein structure, we examined the placement of amino acid differences fixed between the Zanzibari and Nigerien *S. haematobium* populations with PyMol v2.2.3[56].

Invadolysin (Smp_127030) expression - We quantified gene expression of Smp_127030 in *S. haematobium* using previously published data from schistosome eggs, and adult worms from both sexes [15] (SRA accessions: SRX3632877, SRX3632879, SRX3632881). Reads were mapped to the *S. haematobium* genome (SchHae_1.0) using hisat v2.1.0[57]. Transcripts were assembled and gene expression was normalized and quantified using Stringtie v1.3.4[58]. Gene coordinates were lifted from the *S. mansoni* assembly to identify which of the Stringtie predicted transcripts was invadolysin (Smp_127030) using progressiveCactus' halLiftover and the whole genome alignment generated (described above).

Given the limited gene expression data available for *S. haematobium*, we examined expression of Smp_127030 and other invadolysin paralogs in *S. mansoni*[17]. We used RNA-seq data from miracidia[59], cercariae[18], schistosomula (3h and 24h old; in vitro transformation[18)], sporocysts (48h old; in vitro transformation; [59)], juveniles (single sex; 18, 21, 28 day old; [60)], and adults (38 day old; separate males and females from mix infections; [60)]. We aligned the data using STAR v2.5.4b[61]. STAR references were prepared using the v7 *Schistosoma mansoni* genome (ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/v7/sequence/Smansoni_v7.fa.gz), the related v7.1 annotation (ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/v7/annotation/Sm_v7.1.gff.gz; [16)] which was converted to the GTF format using gffread from cufflinks v2.2.1[62] and either an overhang (-sjdbOverhang) of 75 (used for schistosomula and cercariae data) or 99 (used for all

other libraries). We used RSEM[63] to compute transcript per million (TPM) counts for each isoform.

Data Availability

Sequence data are deposited in NCBI's under BioProject accessions PRJNA508633 and PRJNA443709. Mitochondrial genomes are available in are deposited in Genbank (MK253567-MK253578).

Code Availability

Scripts, code, and environmental files are available at https://github.com/nealplatt/sH_hybridization (release v1.0; doi: 10.5281/zenodo.2536390).

References

1. Enciso-Romero, J. *et al.* Evolution of novel mimicry rings facilitated by adaptive introgression in tropical butterflies. *Molecular ecology* **26**, 5160-5172 (2017).
2. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194 (2014).
3. Jones, M.R. *et al.* Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science* **360**, 1355-1358 (2018).
4. Taylor, S.A. & Larson, E.L. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology & Evolution* **3**, 170-177 (2019).
5. King, K.C., Stelkens, R.B., Webster, J.P., Smith, D.F. & Brockhurst, M.A. Hybridization in parasites: consequences for adaptive evolution, pathogenesis, and public health in a changing world. *PLoS pathogens* **11**, e1005098 (2015).
6. Leger, E. & Webster, J.P. Hybridizations within the genus Schistosoma: implications for evolution, epidemiology and control. *Parasitology* **144**, 65-80 (2017).
7. Boissier, J. *et al.* Outbreak of urogenital schistosomiasis in Corsica (France): an epidemiological case study. *The Lancet Infectious Diseases* **16**, 971-979 (2016).
8. Huyse, T., Van den Broeck, F., Hellemans, B., Volckaert, F. & Polman, K. Hybridisation between the two major African schistosome species of humans. *International journal for parasitology* **43**, 687-689 (2013).

9. Huyse, T. *et al.* Bidirectional introgressive hybridization between a cattle and human schistosome species. *PLoS Pathogens* **5**, e1000571 (2009).
10. Sene-Wade, M., Marchand, B., Rollinson, D. & Webster, B.L. Urogenital schistosomiasis and hybridization between Schistosoma haematobium and Schistosoma bovis in adults living in Richard-Toll, Senegal. *Parasitology*, 1-4 (2018).
11. Webster, B.L., Diaw, O.T., Seye, M.M., Webster, J.P. & Rollinson, D. Introgressive hybridization of Schistosoma haematobium group species in Senegal: species barrier break down between ruminant and human schistosomes. *PLoS neglected tropical diseases* **7**, e2110 (2013).
12. Le Clec'h, W. *et al.* Whole genome amplification and exome sequencing of archived schistosome miracidia. *Parasitology*, 1-9 (2018).
13. Bi, K. *et al.* Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC genomics* **13**, 403 (2012).
14. Chevalier, F.D., Valentim, C.L., LoVerde, P.T. & Anderson, T.J. Efficient linkage mapping using exome capture and extreme QTL in schistosome parasites. *BMC genomics* **15**, 617 (2014).
15. Young, N.D. *et al.* Whole-genome sequence of Schistosoma haematobium. *Nature genetics* **44**, 221 (2012).
16. Coghlan, A. *et al.* Comparative genomics of the major parasitic worms. *Nature genetics* **51**, 163-174 (2019).
17. Berriman, M. *et al.* The genome of the blood fluke Schistosoma mansoni. *Nature* **460**, 352 (2009).
18. Protasio, A.V. *et al.* A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni. *PLoS neglected tropical diseases* **6**, e1455 (2012).
19. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* (2009).
20. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489 (2009).
21. Patterson, N.J. *et al.* Ancient admixture in human history. *Genetics*, genetics. 112.145037 (2012).

22. Brisbin, A. *et al.* PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human biology* **84**, 343 (2012).
23. Schumer, M., Cui, R., Powell, D.L., Rosenthal, G.G. & Andolfatto, P. Ancient hybridization and genomic stabilization in a swordtail fish. *Molecular ecology* **25**, 2661-2679 (2016).
24. Chabasse, D., Bertrand, G., Leroux, J., Gauthey, N. & Hocquet, P. Developmental bilharziasis caused by *Schistosoma mansoni* discovered 37 years after infestation. *Bulletin de la Societe de pathologie exotique et de ses filiales* **78**, 643-647 (1985).
25. King, C.H., Muchiri, E.M. & Ouma, J.H. Evidence against rapid emergence of praziquantel resistance in Schistosoma haematobium, Kenya. *Emerging Infectious Diseases* **6**, 585 (2000).
26. Foll, M. & Gaggiotti, O.E. A genome scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* (2008).
27. Sabeti, P.C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913 (2007).
28. Smith, J., Coop, G., Stephens, M. & Novembre, J. Estimating time to the common ancestor for a beneficial allele. *Molecular biology and evolution* **35**, 1003-1017 (2018).
29. Criscione, C.D., Valentim, C.L., Hirai, H., LoVerde, P.T. & Anderson, T.J. Genomic linkage map of the human blood fluke Schistosoma mansoni. *Genome biology* **10**, R71 (2009).
30. Russell, D. & Wilhelm, H. The involvement of the major surface glycoprotein (gp63) of Leishmania promastigotes in attachment to macrophages. *The Journal of Immunology* **136**, 2613-2620 (1986).
31. Yao, C., Donelson, J.E. & Wilson, M.E. The major surface protease (MSP or GP63) of *Leishmania* sp. Biosynthesis, regulation of expression, and function. *Molecular and biochemical parasitology* **132**, 1-16 (2003).
32. Liu, S. *et al.* Expression profile of the Schistosoma japonicum degradome reveals differential protease expression patterns and potential anti-schistosomal intervention targets. *PLoS computational biology* **10**, e1003856 (2014).

33. Curwen, R.S., Ashton, P.D., Sundaralingam, S. & Wilson, R.A. Identification of novel proteases and immunomodulators in the secretions of schistosome cercariae that facilitate host entry. *Molecular & Cellular Proteomics* **5**, 835-844 (2006).

34. Hambrook, J.R., Kaboré, A.L., Pila, E.A. & Hanington, P.C. A metalloprotease produced by larval Schistosoma mansoni facilitates infection establishment and maintenance in the snail host by interfering with immune cell function. *PLOS Pathogens* **14**, e1007393 (2018).

35. Storz, J.F., Opazo, J.C. & Hoffmann, F.G. Gene duplication, genome duplication, and the functional diversification of vertebrate globins. *Molecular phylogenetics and evolution* **66**, 469-478 (2013).

36. Emery, A.M., Allan, F.E., Rabone, M.E., Rollinson, D.J.P. & vectors. Schistosomiasis collection at NHM (SCAN). **5**, 185 (2012).

37. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754-1760 (2009).

39. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

40. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* (2010).

41. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

42. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

43. Paten, B. *et al.* Cactus: Algorithms for genome multiple sequence alignment. *Genome research* (2011).

44. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559-575 (2007).

45. Browning, B.L. & Browning, S.R. Genotype imputation with millions of reference samples. *The American Journal of Human Genetics* **98**, 116-126 (2016).

46. Lockyer, A. *et al.* The phylogeny of the Schistosomatidae based on three genes with emphasis on the interrelationships of Schistosoma Weinland, 1858. *Parasitology* **126**, 203-224 (2003).
47. Consortium, A.g.G. Genetic diversity of the African malaria vector Anopheles gambiae. *Nature* **552**, 96 (2017).
48. Gautier, M. & Vitalis, R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**, 1176-1177 (2012).
49. Finn, R.D. *et al.* Pfam: the protein families database. *Nucleic acids research* **42**, D222-D230 (2013).
50. Rawlings, N.D. & Barrett, A.J. [13] Evolutionary families of metallopeptidases. in *Methods in enzymology*, Vol. 248 183-228 (Elsevier, 1995).
51. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797 (2004).
52. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
53. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular biology and evolution* **33**, 1870-1874 (2016).
54. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. & Sternberg, M.J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols* **10**, 845 (2015).
55. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics* **10**, 168 (2009).
56. DeLano, W.L. PyMOL. (2002).
57. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nature methods* **12**, 357 (2015).
58. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290 (2015).
59. Wang, B., Collins III, J.J. & Newmark, P.A. Functional genomic characterization of neoblast-like stem cells in larval Schistosoma mansoni. *Elife* **2**, e00768 (2013).
60. Protasio, A.V. *et al.* MiR-277/4989 regulate transcriptional landscape during juvenile to adult transition in the parasitic helminth Schistosoma mansoni. *PLoS neglected tropical diseases* **11**, e0005559 (2017).

61. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
62. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511 (2010).
63. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323 (2011).

Acknowledgements

Sandy Smith, Richard Polich and Roy Garcia (Texas Biomedical Research Institute) provided research support. Steffi Knopp (Swiss Tropical and Public Health Institute) Khalfan Mohammed, Mtumweni Ali Muhsin (Helminth Control Laboratory, Unguja) and Mr. Haji Khatib Faki (Public Health Laboratory, Pemba) for their assistance in the collection of the Zanzibar samples. Matthew Berriman and collegues provided genome sequence data prior to publication. Muriel Rabone for mantiaining the SCAN database. This research was funded by the National Institute of Allergy and Infectious Diseases (R01 AI097576-01; 5R21AI096277-01), the National Center for Advancing Translational Sciences (UL1TR001120), Texas Biomedical Research Institute Forum (award 0467), the National Center for Research Resources (C06 RR013556; RR017515), the Wellcome Trust (104958/Z/14/Z), the Gates Foundation coordinated by the University of Georgia Research Foundation Inc (RR374-053/5054146; RR374-053/4785426), and a ZELS research grant (combined BBSRC, MRC, ESRC, NERC, DSTL and DFID: BB/L018985/1) . WL was supported by a Cowles Fellowship.

Author contributions

Study Design (T.J.C.A., J.P.W, D.R., A.M.E., B.L.W.), Formal Analysis (R.N.P.II, T.J.C.A., W.L., F.D.C), writing – original draft (R.N.P.II, T.J.C.A); writing – review and editing (M.M.W, W.L, F.D.C, B.L.W, J.P.W, D.R, A.M.E); sampling (J.P.W, D.R., F.A., B.L.W, A.G, F.K, S.M.A, A.M.E).

Competing Interests

The authors declare no competing interests.

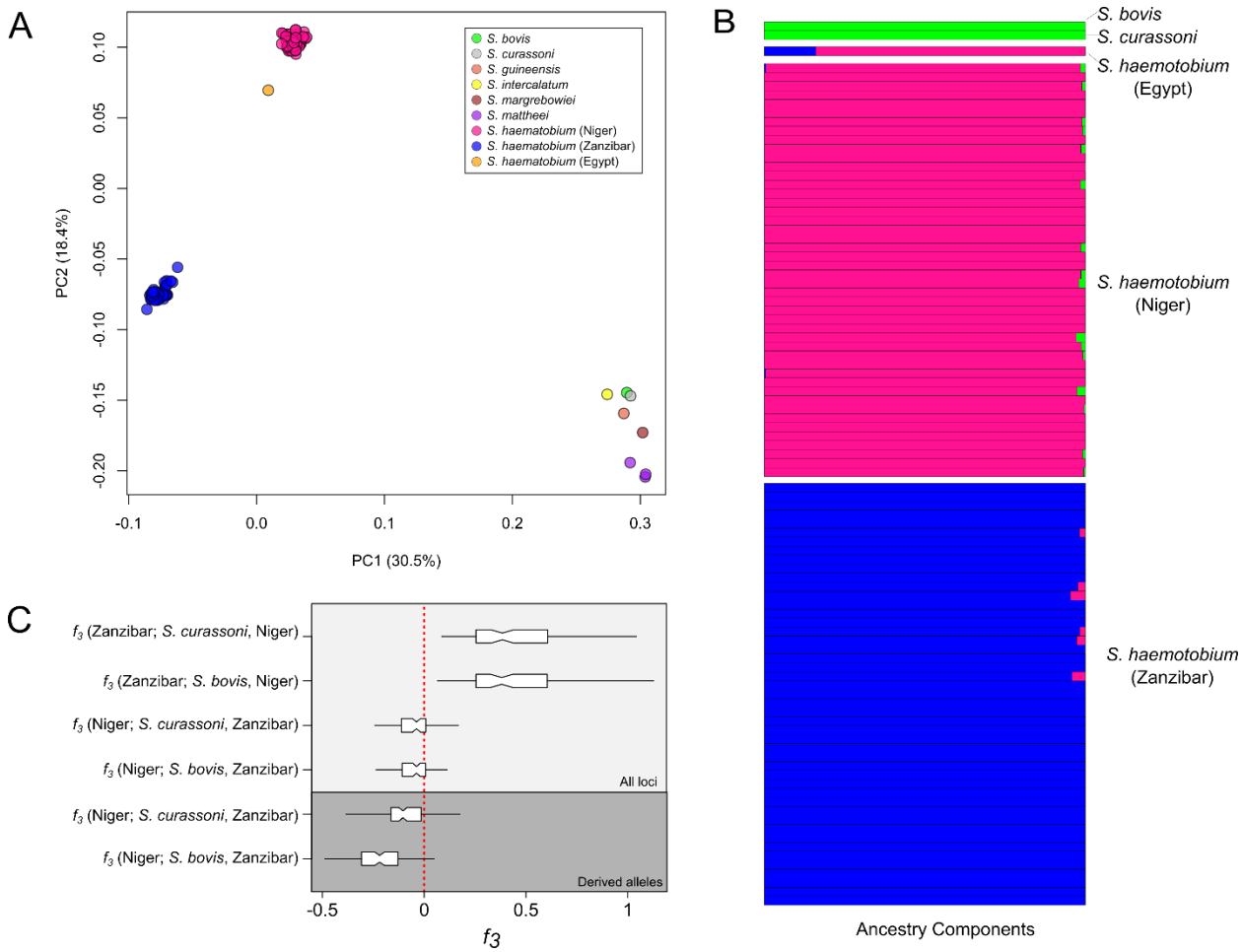

Fig. 1. Population structure in *Schistosoma haematobium*. (A) PCA plot shows clear distinction between the two *S. haematobium* populations and the rest of the species examined: we see no evidence for recent hybridization (B) Quantification of population structure with ADMIXTURE showed low levels of admixture between *S. haematobium* populations and *S. bovis*/*curassoni*. (C) The three-population statistic (f3) was used to formally test admixture between each of the *S. haematobium* populations, *S. bovis*, and *S. curassoni*. When testing f3 (test; A, B), a negative result indicates that the test group is an admixed population from A and B. To differentiate between introgression from *S. bovis* and *curassoni* in Nigerien *S. haematobium*) we recalculated f3 using only derived alleles from *S. bovis* or *curassoni*. These results rule out contemporary hybridization between populations of *S. haematobium* and *S. bovis* or *curassoni*, and indicate that low levels of introgression are restricted to the Nigerien *S. haematobium* population.

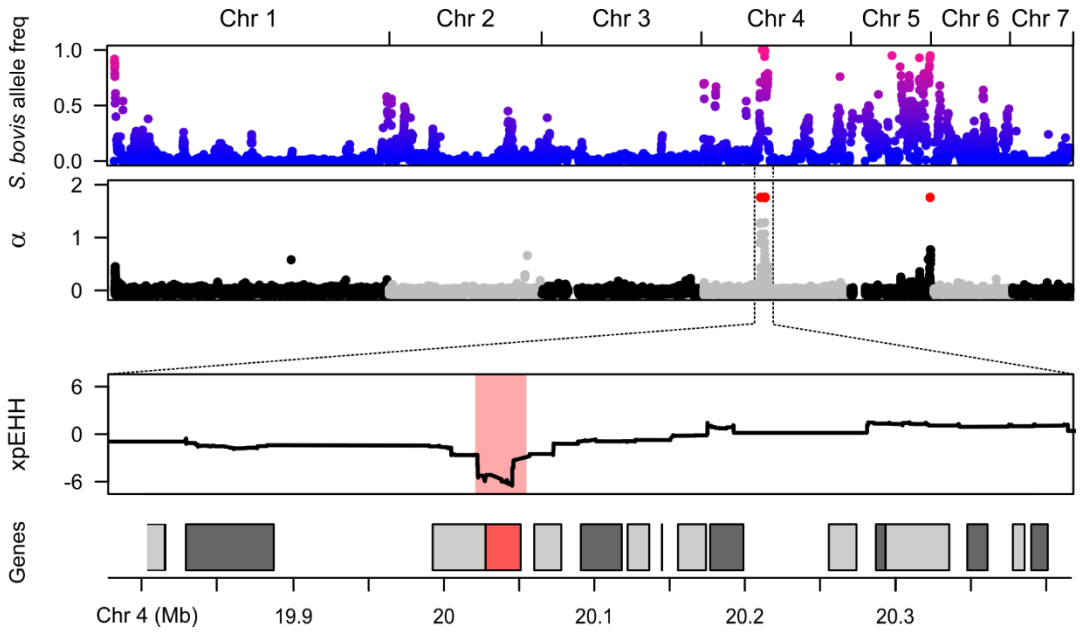

Fig. 2. Selection on introgressed alleles in Nigerien *S. haematobium*. We aligned plots showing (a) proportion of *S. bovis* ancestry (b) genetic differentiation (FST) between Niger and Zanzibari samples (c) the α statistic from BayeScan and (d) xpEHH. S. bovis alleles have reached high frequencies in the Nigerien *S. haematobium* population. Introgressed alleles under directional selection were identified using allele frequency differences (α; alpha) between populations and using regions of extended homozygosity (xpEHH). Each test identified multiple regions under selection, but only one region was identified by both approaches. This region on Chr 4 spanned a single gene (Smp_120703, invadolysin) at which *S. bovis* alleles are approaching fixation in Nigerien *S. haematobium*.