Global diversity of soil-transmitted helminths reveals population-biased genetic variation that impacts diagnostic targets

Marina Papaiakovou[1+], Andrea Waeschenbach[2], Olumide Ajibola[3], Sitara SR Ajjampur[4,5], Roy M Anderson[6], Robin Bailey[5,7], Jade Benjamin-Chung[8,9], Maria Cambra-Pellejà[10,11,12], Nicolas R Caro[13], David Chaima[14,5], Ruben O Cimino[13], Piet Cools[15], Anélsio Cossa[16], Julia Dunn[6], Sean Galagan[5,17], Javier Gandasegui[10,18], Berta Grau-Pujol[10,19,20], Emma L Houlder[21], Moudachirou Ibikounlé[22], Timothy P Jenkins[23], Khumbo Kalua[5,24], Eyrun F Kjetland[25,26], Alejandro J Krolewiecki[13], Bruno Levecke[15], Adrian Luty[5,27], Andrew MacDonald[28], Inácio Mandomando[10,16,29,30], Malathi Manuel[4,5], Maria Martínez-Valladares[31], Rojelio Mejia[32], Zeleke Mekonnen[33], Augusto Messa Jr.[16], Harriet Mpairwe[34], Osvaldo Muchisse[16], Jose Muñoz[10,35,36], Pauline Mwinzi[37,38], Valdemiro Novela[16], Maurice R Odiere[37], Charfudin Sacoor[16], Judd L Walson[5,39], Steven A Williams[40], Stefan Witek-McManus[5,41], D. Timothy J. Littlewood[2*], Cinzia Cantacessi[1*] and Stephen R Doyle[18+].

1. Department of Veterinary Medicine, University of Cambridge, Cambridge, UK
2. Natural History Museum, Cromwell Road, London, UK
3. Department of Biochemistry, Nigerian Institute of Medical Research, Yaba, Lagos
4. The Wellcome Trust Research Laboratory, Division of Gastrointestinal Sciences, Christian Medical College Vellore, Vellore, Tamil Nadu, India
5. The DeWorm3 Project, University of Washington, Seattle, WA, USA
6. Department of Infectious Disease Epidemiology, School of Public Health, Faculty of Medicine, Imperial College London, Wood Lane, White City, London, UK
7. Clinical Research Department, Faculty of Infectious and Tropical Disease, London School of Hygiene and Tropical Medicine, London, UK
8. Department of Epidemiology and Population Health, Stanford University, Stanford, CA, USA
9. Chan Zuckerberg Biohub, San Francisco, CA, USA

10. Barcelona Institute for Global Health (ISGlobal), Hospital Clínic, University of Barcelona, Barcelona, Spain
11. GraphenicaLab SL, Barcelona, Spain
12. Facultat de Medicina i Ciències de la Salut, Universitat de Barcelona, Barcelona, Spain
13. Instituto de Investigaciones de Enfermedades Tropicales (IIET-CONICET), Sede Regional Orán, Universidad Nacional de Salta, Orán, Argentina
14. Department of Microbiology, Kamuzu University of Health Sciences, Blantyre, Malawi
15. Department of Translational Physiology, Infectiology and Public Health, Ghent University, Merelbeke, Belgium
16. Centro de Investigação em Saúde de Manhiça (CISM), Maputo, Mozambique
17. Department of Global Health, University of Washington, Seattle, WA, USA
18. Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK
19. Health Manhiça Research Center (CISM), Maputo, Mozambique
20. Mundo Sano Foundation, Buenos Aires, Argentina
21. Leiden University Center for Infectious Diseases, Leiden University Medical Center, Leiden, Netherlands
22. Centre de Recherche pour la lutte contre les Maladies Infectieuses Tropicales (CReMIT/TIDRC), Université d'Abomey-Calavi, Bénin
23. Department of Biotechnology and Biomedicine, Technical University of Denmark, Kongens Lyngby, Denmark
24. Blantyre Institute for Community Outreach, Lions Sight First Eye Hospital, Blantyre, Malawi
25. Norwegian Centre for Imported and Tropical Diseases, Department of Infectious Diseases Ullevaal, Oslo University Hospital, Oslo, Norway
26. Discipline of Public Health Medicine, Nelson R Mandela School of Medicine, College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa
27. Université Paris Cité, IRD, MERIT,Paris, France
28. Lydia Becker Institute of Immunology and Inflammation, University of Manchester, Manchester, UK
29. Global Health and Tropical Medicine, GHTM, Associate Laboratory in Translation and Innovation Towards Global Health, LA-REAL, Instituto de Higiene e Medicina Tropical, IHMT, Universidade NOVA de Lisboa, UNL, Lisboa, Portugal
30. Instituto Nacional de Saúde (INS), Mozambique
31. Instituto de Ganadería de Montaña, CSIC, Universidad de León, Grulleros, León, Spain

32. Departments of Pediatrics and Medicine, National School of Tropical Medicine, Baylor College of Medicine, Houston, TX, USA
33. Department of Medical Laboratory Sciences, Institute of Health, Jimma University, Jimma, Ethiopia
34. MRC/UVRI and LSHTM Uganda Research Unit, Entebbe, Uganda
35. Centre for Global Health Research, Kenya Medical Research Institute, Kisumu, Kenya
36. International Health Department, Hospital Clinic de Barcelona, Barcelona, Spain
37. Universitat de Barcelona, Barcelona, Spain
38. Expanded Special Project for Elimination of NTDs, WHO Regional Office for Africa, Brazzaville, Republic of Congo
39. Departments of International Health, Medicine and Pediatrics (Infectious Disease), Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
40. Smith College, Northampton, MA, USA
41. Department of Disease Control, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK

* These authors contributed equally

[+]**Corresponding authors**

Marina Papaiakovou (mpapaiakovou@gmail.com), Stephen R. Doyle (stephen.doyle@sanger.ac.uk)

Abstract

Soil-transmitted helminths (STHs) are intestinal pathogens affecting over a billion people worldwide. STH control relies on microsopy-based diagnostics to monitor parasite prevalence and enable post-treatment surveillance, however, molecular diagnostics are rapidly being developed due to increased sensitivity, particularly in low-STH-prevalence settings. The genetic diversity of helminths and its potential impact on molecular diagnostics remains unclear. Using low-coverage genome sequencing, we assessed the presence of STHs in worm, faecal and purified egg samples from 27 countries, identifying differences in the genetic connectivity and diversity of STH-positive samples across regions and cryptic diversity between closely related human- and pig-infective species. We defined substancial copy number and sequence variants in current diagnostic targets and validated the impact of genetic variants on qPCR diagnostics using *in vitro* assays. Our study provides insights into the diversity and genomic epidemiology of STHs, highlighting challenges and opportunities for developing molecular diagnostics needed to support STH control efforts.

Main

Soil-transmitted helminths (STHs), which include the giant roundworm (*Ascaris lumbricoides),* threadworm *(Strongyloides stercoralis)*, whipworms (*Trichuris trichiura)*, and hookworms (*Necator americanus* and *Ancylostoma duodenale)*, collectively infect nearly a quarter of the world's human population. Infections are established from the consumption of infective eggs or penetration of the skin by infective larvae followed by migration to the hosts' intestines, where STHs sexually mature, reproduce and lay eggs that are passed in the faeces[1,2]. *Strongyloides* species are unique in that their eggs hatch within the host, and larvae pass into the faeces, some of which can develop directly into infectious larvae and autoinfect the host[3]. Light intensity STH infections are often asymptomatic, but moderate-to-heavy intensity infections can cause significant morbidity[4], including anaemia, malaise and weakness, intestinal obstruction, malnutrition, impaired growth and developmental delay[5]. Control of STHs relies almost exclusively on annual or biannual mass drug administration (MDA) of anthelmintics to at-risk populations (preschool and school-aged children (SAC) as well as women of reproductive age) or to entire communities to reduce morbidity and transmission[6]; along with health education and the improvement of water, sanitation and hygiene conditions [7]. This control strategy is central to the World Health Organization's (WHO) 2030 Roadmap for STH, which aims to achieve and maintain the elimination of morbidity caused by STHs as a public health problem, and to establish an efficient strongyloidiasis control programme in SAC by 2030[8].

A critical aspect of the WHO's control strategy for STHs is the use of diagnostics to detect and monitor infections, from establishing baseline prevalence to confirming elimination as a public health problem once MDA campaigns have concluded[9]. The WHO currently recommends microscopy-based diagnostic methods, which are particularly suited to diagnose moderate-to-heavy intensity infection burdens due to their simplicity and relatively low cost. The Kato-Katz light microscopy technique is used to detect helminth eggs in smears of faeces[10]. However, its sensitivity is significantly reduced when infection burdens are low[11], leading to variability in parasite detection. Furthermore, this approach is ineffective for diagnosing *Strongyloides* infections due to its intermittent larvae excretion and low parasite load[12]. DNA-based diagnostics, such as quantitative polymerase chain reaction (qPCR) assays, are beginning to offer an alternative to conventional microscopy-based STH diagnosis methods due to the increase in their sensitivity and specificity in low prevalence settings[13] and their potential for application during post-deworming surveillance where microscopy approaches become less efficient owing to the reduced intensity of infection[14–18]. Current qPCR assays primarily were

developed and validated using a single or limited number of geographically restricted isolates[19–23]. However it is increasingly recognised that parasitic worms of humans[24–29] and animals[30–32] are genetically diverse and distributed within genetically structured populations at local and global scales. These genetic variations can occur in sequences targeted by molecular methods, and therefore, the variation within single species that differentiates STH populations may also affect the sensitivity and specificity of the diagnostic tests in different settings. Yet to date, there are only a few studies on the population genetics of STHs[33], limiting our understanding of the extent of genetic variation that might interfere with molecular diagnostic methods.

To understand the population genetics of STHs and the impact of genetic variation on molecular diagnostics, we analysed low-coverage whole genome and metagenomic sequencing of worm (n = 128), faecal (n = 842) and purified egg (n = 30) samples samples spanning 27 STH-endemic countries from around the world. Analyses of these data revealed the abundance and breadth of single- and mixed-species STH infections, the genetic connectivity of STH populations at different geographic scales, and significant genetic variation at sites targeted by diagnostics, both within genomes and between countries. Using *in vitro* assays, we assessed the impact of genetic variants on qPCR diagnostics. These data will support molecular epidemiological and emerging genomic efforts to understand STH transmission patterns and help validate current and future diagnostic tools to support the control and elimination of STHs.

Results

Detection of single- and co-infections of STH from faecal, egg and worm isolates

To understand the genetic diversity and differentiation of STH populations worldwide, we analysed data from 1000 DNA samples extracted from individual worms (n = 128) as well as faecal (n = 842) and egg samples (n = 30), from 27 countries across all continents (Fig. 1 and Supplementary Table 1). Sequencing reads were mapped to eight mitochondrial and nuclear reference genomes of STHs and other gastrointestinal helminths reported to infect humans (Supplementary Table 2). Out of 842 faecal samples, 175 were positive for at least one helminth species (threshold of a minimum of 10 reads / million reads mapped; Supplementary Fig. 1a) spanning 14 different countries (Fig. 1a, c and Supplementary Fig. 2a). Reads mapping to *A. lumbricoides* were the most abundant (96 single-infections and 27 co-infections containing two or more species), which might be expected given that the high fecundity of *Ascaris* worms (females deposit >200,000 eggs per day) would result in a larger proportion of *Ascaris* DNA being sequenced. The next largest number of infections was caused by *N. americanus* (35 single infections and 13 co-infections) and then *T. trichiura* (six single infections and 15 co-infections). We also assessed other helminth species that can be present in faecal samples, mapping reads to *Schistosoma mansoni* (eight single infections and one co-infection), *Strongyloides stercoralis* (one co-infection), *Enterobius vermicularis* (two co-infections) and *Ancylostoma ceylanicum* (one single infection and two co-infections). Among the 158 samples derived from worms or pools of eggs, 154 were helminth-positive, spanning 15 different countries (Figs. 1b,d and Supplementary Fig. 1b). *A. lumbricoides* was the most abundant helminth present in the cohort tested (71 single infections and 20 co-infections), followed by *T. trichiura* (26 single infections and 22 co-infections), *S. stercoralis* (16 single infections), *N. americanus* (five single infections, nine co-infections), *S. mansoni* (11 single infections, one co-infection), and one co-infection each, by *E. vermicularis* and *Ancylostoma duodenale*.

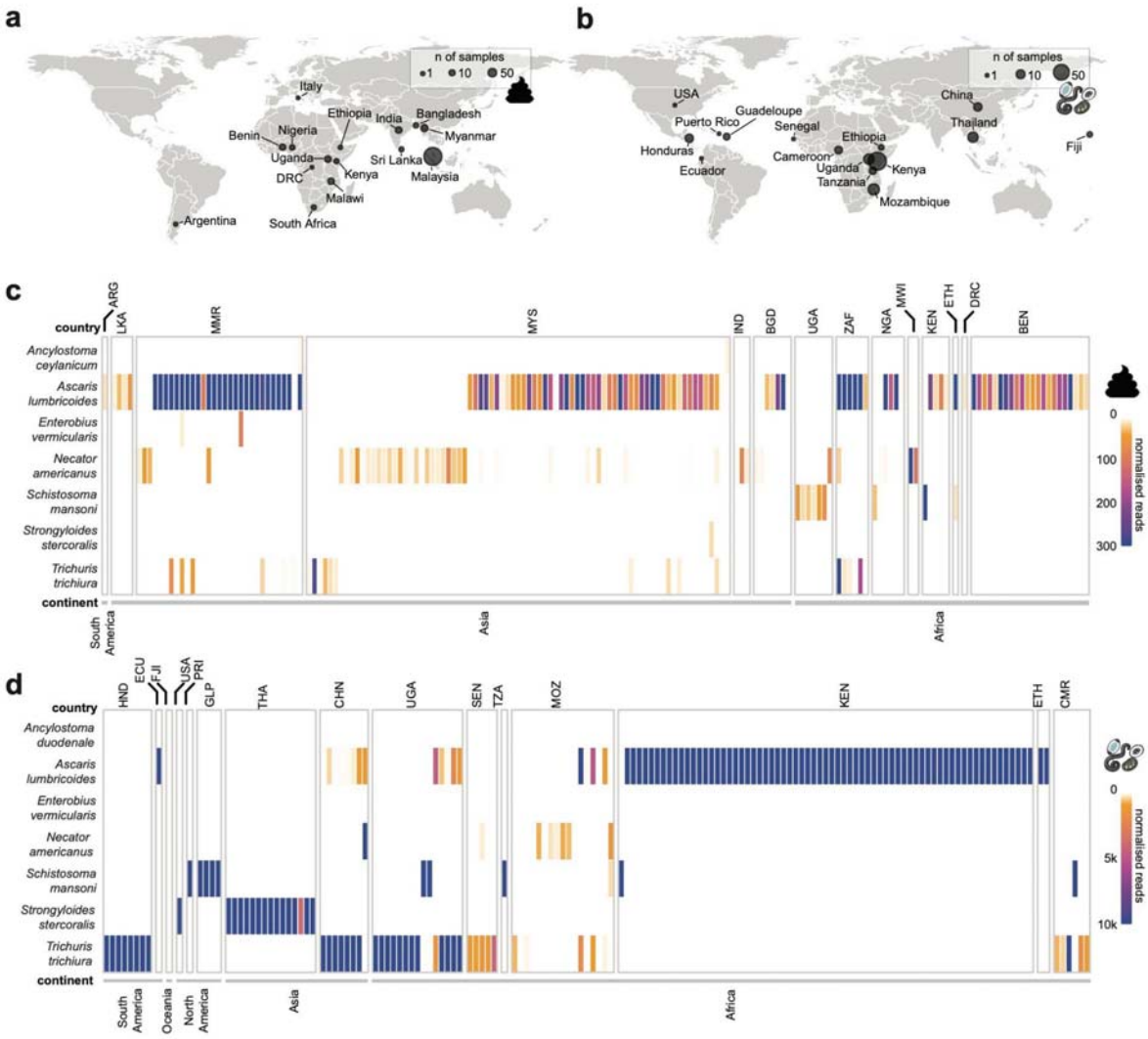


Fig. 1: The geographic distribution of gastrointestinal helminths detected among worm/egg isolates and faecal samples analysed in this study.

a,b, World maps show the country of origin of all the samples used in the study, which include (**a**) faecal samples (n = 842) spanning 15 countries and (**b**) worm and concentrated egg isolates (n = 158) spanning 15 countries. The point size on the maps shows the number of samples from each location. The faecal sample icon indicates faecal samples, and adult worm/egg shapes indicate samples from adult worms and/or concentrated worm eggs. **c,d,** Heatmaps showing the prevalence of single and mixed-species infections of parasites found in (c) faecal samples or (d) worm/egg isolates, respectively. Colours reflect read counts, normalised by the total number of reads per sample per genome size, to obtain 'reads mapped per million reads per Mb'. A minimum normalised coverage threshold of 10 reads was applied, resulting in 329 samples from 27 countries reported as positive for at least one parasite. Sample site abbreviations for both faecal and worm data are as follows: ARG = Argentina; BEN = Benin; BGD =

Bangladesh; CHN = China; CMR = Cameroon; DRC = Democratic Republic of the Congo; ECU = Ecuador; ETH = Ethiopia; FJI = Fiji; GLP = Guadeloupe; HND = Honduras; IND = India; KEN = Kenya; LKA = Sri Lanka; MOZ = Mozambique; MMR = Myanmar; MWI = Malawi; MYS = Malaysia; NGA = Nigeria; PRI = Puerto Rico; SEN = Senegal; THA = Thailand; TZA = Tanzania; UGA = Uganda; USA = United States of America; ZAF = South Africa.

Mitochondrial genetic diversity within and between populations of *Ascaris* spp. and *Trichuris trichiura*

To estimate the genetic diversity within and between populations of each STH species, we focused our analyses on samples with a minimum 5X sequencing coverage to support variant calling. From the sample cohort samples, we retained 108 *A. lumbricoides*-positive samples from 10 countries, 35 *T. trichiura*-positive samples from seven countries, 16 *S. stercoralis*-positive samples from two countries, 12 *S. mansoni*-positive samples from seven countries, and seven *N. americanus*-positive samples from two countries. Focusing on *A. lumbricoides and T. trichiura,* we identified 2,054 robust mitochondrial single nucleotide polymorphisms (SNPs) (Supplementary Table 3) for population genetic analyses: 558 SNPs in *A. lumbricoides* positive samples (n = 88 samples) and 1,496 SNPs (n = 30 samples) in samples positive for *T. trichiura*. Considering that no SNPs from *N. americanus* remained after stringent filtering (Supplementary Table 3)*, S. stercoralis*-positive samples represented only two countries (15 samples from Thailand and one from the USA), and the population genetic analysis of *S. mansoni* datasets has already been explored[34], these helminths were excluded from further analysis.

Initial analyses of *A. lumbricoides* genetic variation estimated by mapping to a Korean mitochondrial reference genome of this species isolated from a human (Accession: NC_016198;[35] showed unusually low between-country and high within-country (specifically, within Kenya) genetic diversity and a lack of country-specific population clustering (Supplementary Figs. 3a,b,c). These clustering patterns could not be attributed to missing variants or coverage biases (Supplementary Fig. 3d). *Ascaris* is a known zoonotic pathogen, and there is ongoing debate as to whether human-infective *A. lumbricoides* and the pig-associated *A. suum* are a single or distinct species[36,37]. To explore potential reference mapping biases, we mapped *Ascaris*-positive samples to two additional *Ascaris* mitochondrial reference genomes, one isolated from a pig (*A. suum* from the US; Accession: NC_001327;[38]) and one from a human (*A. lumbricoides* from Tanzania; Accession: KY045802). Competitive mapping to

the three references showed a bias for the *A. suum* reference rather than either of the *A. lumbricoides* references (Supplementary Fig. 4a). Only six samples mapped preferentially to the Korean *A. lumbricoides* reference (Supplementary Fig. 4b). The remaining samples fit two *A. suum*-like profiles: (i) majority *A. suum* with partial Tanzania *A. lumbricoides*, and (ii) majority *A. suum* only. The first was primarily driven by the Kenyan samples, previously described as displaying *A. suum*-like variation[37]; however, we also observed the same pattern in approximately half the samples from Myanmar. The remaining countries fit the second profile, with minimal *A. lumbricoides* variation present. The six Kenyan outliers that mapped preferentially to the *A. lumbricoide*s reference from Korea were excluded from downstream population analyses. Remapping exclusively to the *A. suum* reference, followed by stringent variant calling, revealed 346 SNPs, which improved the genetic resolution of the remaining 81 samples (Supplementary Figs. 5a,b,c).

To measure genetic variation between and within populations, we assessed pairwise genetic differentiation (D_{XY}), nucleotide diversity (π), and the frequency of alleles ($freq_{allele}$) (for pooled populations) or frequency of SNPs ($freq_{SNP}$) (for individual worms). D_{XY} showed contrasting patterns between *A. lumbricoides* and *T. trichiura* (Fig. 2a). For *T. trichiura,* we found evidence of low genetic differentiation between sub-populations of parasites within a given country and that populations from geographically close countries were more genetically similar than from more geographically distant countries (Fig. 2a, Supplementary Figs. 6a,b; meanD_{XY}[between] / meanD_{XY}[within] = 5.59X and >33X for worm and egg, respectively), with genetically distinct populations (Supplementary Figs. 6c,d) consistent with previous studies of *Trichuris* genetic variation[24]. By contrast, *A. lumbricoides* showed low levels of differentiation comparing within versus between country D_{XY} (meanD_{XY}[between] / meanD_{XY}[within] = 2.48X and 1.47X for worm and egg, respectively) and significantly lower levels of genetic differentiation between populations despite geographic distance, even between populations from Africa and Asia (Fig. 2a, Supplementary Figs. 5a,b,c,d). The contrasting patterns between species are likely driven by consistent differences in the patterns of genetic variation within populations as measured by nucleotide diversity (Fig. 2b) and variant frequency (Fig. 2c) in both worm and egg samples. In almost all cases, nucleotide diversity and variant frequency were higher in *T. trichiura* than in *A. lumbricoides* populations. This pattern of variation in *A. lumbricoides* is consistent with that of Easton *et al.,*[37], where very few variants were found to differentiate haplotypes from reference genomes sampled from around the world. Collectively, these data suggest that *Ascaris* has more recently and more rapidly spread around the world, likely supported by relatively higher

reproductive output and zoonotic transmission that may have been exacerbated by modern livestock transport, homogenising geographically distant populations[30]. Further targeted sampling of *Ascaris* from humans and pigs living within proximity or cohabiting in endemic regions will shed further light on zoonotic transmission and its contribution to parasite radiation.

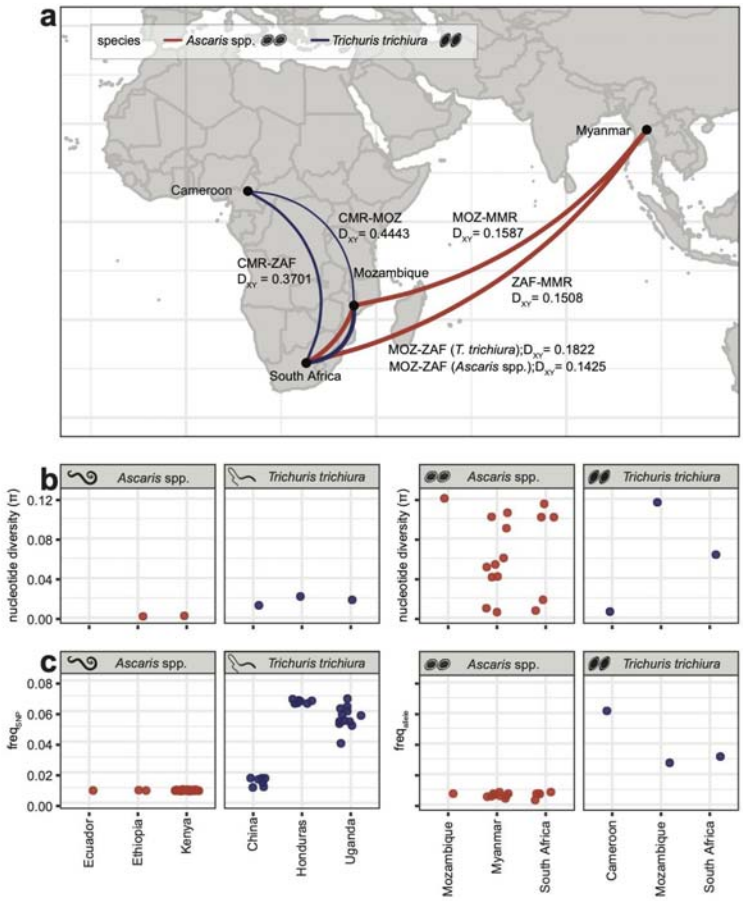


Fig. 2: Comparison of genetic diversity within and between populations of *Trichuris trichiura* and *Ascaris* spp.

a, Relative genetic relationships between populations for pooled samples based on pairwise estimates of mitochondrial genome diversity (D_{XY}). Colour indication: blue = *Trichuris trichiura*, red = *Ascaris* spp. The thickness of the line connecting countries reflects the degree of similarity between mitochondrial genomes; thicker and thinner lines represent stronger and weaker genetic similarity, respectively. The mean D_{XY} per country combination was calculated, and the line size was plotted as '1- mean D_{XY} per country combination'. **b,** Comparison of nucleotide diversity (π) in mitochondrial genomes amongst individual worms (left) and pools of eggs from faecal samples (right) (n = 16 for *Ascaris* spp., n = 3 for *T. trichiura*) per population. **c,** Comparison of the variant frequencies as SNPs in the individuals' populations (n = 28 samples for *T. trichiura*, n = 65 for *Ascaris* spp.) and alleles in the pools (n = 16 for *Ascaris* spp., n = 3 for *T. trichiura*). The frequencies of SNPs and alleles were calculated to enable the comparison of genetic variation amongst both individual worms and pools of eggs. For Ecuador, the calculation of nucleotide diversity was infeasible due to the availability of a single sample. Country codes are as follows: CMR = Cameroon; MMR = Myanmar; MOZ = Mozambique; ZAF = South Africa.

Nuclear diagnostic markers vary substantially in number and sequence within and between STH species

We next sought to understand the degree to which nuclear diagnostic markers, specifically those used as targets for qPCR, vary in sequence and abundance in the genomes of STHs. Characterisation of the location of repeat arrangements in the *A. lumbricoides*, *N. americanus,* and *T. trichiura* reference genome assemblies (Supplementary Figs. 7a,b,c, respectively) revealed coordinates for (i) 125 germline tandem repeat targets (spread across 11 scaffolds) and six ITS targets (across two scaffolds) for *A. lumbricoides*, (ii) 302 tandem repeat targets (across 11 scaffolds) and 15 18S targets (in one scaffold) for *T. trichiura*, and (iii) 336 tandem repeat targets (across eight scaffolds) and two ITS targets (in one scaffold) for *N. americanus* (for marker details, see Supplementary Table 4). A pairwise comparison of all of those repeats within each species highlighted substantial sequence diversity of repeats within each species (Fig. 3 a,b,c), despite all repeats being identified using a single canonical repeat target sequence as a query. The sequence diversity among repeats suggests that not all sequences are suitable qPCR targets; only 80/125 (*A. lumbricoides*), 221/302 (*T. trichiura*), and 335/336 (*N. americanus*) sequences included the complete primer and probe binding sites necessary for amplification and detection, respectively.

Our understanding of the repeat arrays in a genome, as observed above, is based on a single reference genome, likely containing technical variation (i.e. collapsed repeats from misassembly) and biological variation (i.e. repeat copy number) compared to the original isolate sequenced. To explore the latter, we estimated the coverage of each repeat per country per species, which was normalised by the coverage of exons from single-copy genes. Overall, *A. lumbricoides* (Fig. 3d; n = 69/80 with coverage > 0) and *N. americanus* (Fig. 3e; n = 49/335 with coverage > 0) contained fewer repeats with more variable coverage relative to *T. trichiura* (Fig. 3f; n = 81/221 with coverage). We attribute this to the difference between short-read assemblies of *A. lumbricoides* and *N. americanus* (prone to collapsed repeats) and the more contiguous long-read assembly of *T. trichiura* in which the repeat arrays are likely to be better assembled. Nonetheless, repeats with coverage varied substantially between countries and repeat types, suggesting that population-specific variation in repeat copy number may play a role in the qPCR-based diagnosis of STHs.

The sensitivity and specificity of qPCR assays targeting repetitive sequences are influenced by whether the target(s) lies within the nuclear or mitochondrial genome. Overall, the relative copy number of nuclear repeats was higher than that of mitochondrial genomes (except for adult *A. lumbricoides,* which had a larger copy number of mitochondrial genomes than nuclear repeats likely due to repeat loss in adults as a consequence of chromosome diminution[39]). These ratios were more consistent across species in eggs than in adults (Supplementary Figs. 8a,b). Considering that, in practice, the DNA of eggs from faeces is the primary target of the qPCR diagnostic for STHs, our data support the underlying hypothesis that nuclear repeats provide a higher degree of sensitivity (relative to mitochondrial genome targets or other single-copy genes), albeit with the caveat of greater variation of nuclear repeats within and between STH species than previously appreciated[40].

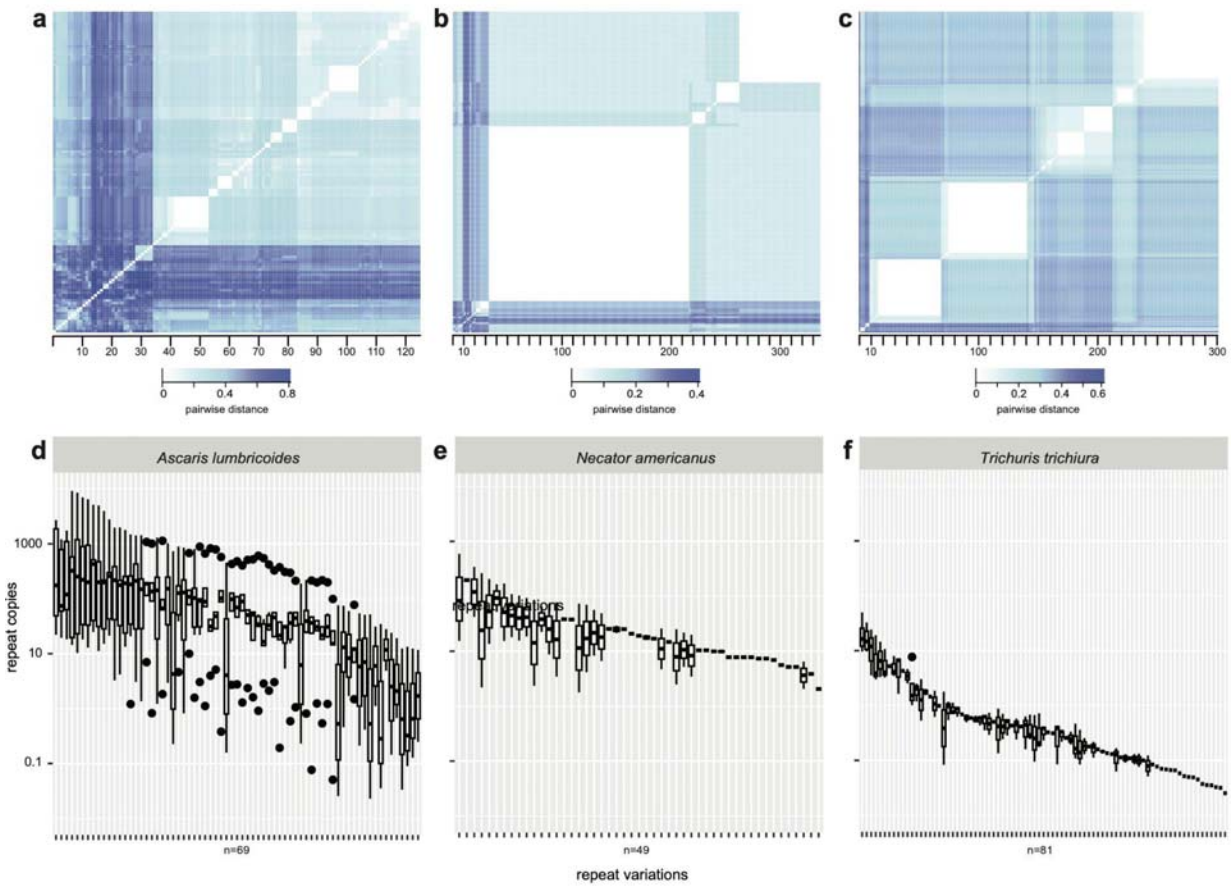


Fig. 3: Genomic diversity and differential coverage of repeat classes used as diagnostic targets.

a,b,c, Repeat diversity was determined by searching the canonical repeat units using *Nucmer* with 90% similarity and 90% coverage against genome assemblies of (**a**) *Ascaris lumbricoides*, (**b**) *Necator americanus* and (**c**) *Trichuris trichiura* as a reference. In **a, b,** and **c,** the heatmaps show the pairwise

distance calculated by the sum of squares of a similarity matrix from ClustalOmega-aligned repeat sequences from each species. In **d, e,** and **f**, Genome coverage per repeat per country was determined by *bedtools multicov* (with minimum overlap 0.51) in merged-by-country BAM files (filtered raw reads > ten reads) against the genome assemblies of (**d**) *A. lumbricoides*, (**e**) *N. americanus,* and (**f**) *T. trichiura*. Coverage is expressed as 'repeat copies' calculated by original repeat coverage divided by the mean-per-country single copy exon coverage. Only repeats containing both forward and reverse primers and probe binding sites were included.

Genetic variation within primer sites can detrimentally affect qPCR amplification of diagnostic targets

Sequence and copy number variation within and between populations prompted us to determine whether individual genetic variants within nuclear repeats, particularly in the primer binding sites, would affect the diagnostic performance of the qPCR assays. Stringent variant filtering of nuclear-mapped data (Supplementary Table 5), keeping only variants on oligo (primer/probe) binding sites, identified in 144 SNPs (n = 68 samples) for *A. lumbricoides*, 508 SNPs (n = 29 samples) for *T. trichiura* and 123 SNPs (n = 4 samples) for *N. americanus*. To test the effect of these SNPs on qPCR assays, we identified repetitive sequences for each species that (i) contained SNPs that varied substantially in frequency between populations and (ii) contained SNPs within five bp from the 3' end of either of the primers that would result in a priming-mismatch. From these variants, we selected three tandem repeats from each of *A. lumbricoides* (Fig. 4a) and *T. trichiura* (Supplementary Fig. 9a) for validation. No variants meeting those criteria were found for *N. americanus*.

To assess whether these variants in the binding sites might impact primer annealing and amplification, we evaluated the impact of these variants in established qPCR assays[19,20]. This was done using serial dilution series of cloned target sequences, each containing either the "wildtype" or "mutant" base (Supplementary Table 6). For *A. lumbricoides*, the presence of the variant in both repeats 1 and 2 (Al_SNP_1 and Al_SNP_2; both A-to-G variants, both in scaffold_3) did not reduce the efficiency of the qPCR assay, whereas, in repeat 3 (Al_SNP_3; T-to-G variant), the efficiency decreased to 83.9% (Fig. 4b). However, the presence of variants in the primer binding sites resulted in >1000-fold loss of amplified product in repeat 2 (Al_SNP_2), followed by >500-fold loss in repeat 1 (Al_SNP_1), and >50-fold loss in repeat 3 (Al_SNP_3) (Fig. 4c). We observed similar patterns for *T. trichiura*. The presence of the SNP in

repeat 1 (Tt_SNP_1; T-to-G variant), repeat 2 (Tt_SNP_2; G-to-C variant) and repeat 3 (Tt_SNP_3; T-to-A variant) prevented any amplification of the target (Supplementary Fig. 9b). However, the variants occurred in regions where the primers and probe had several mismatches with both the wildtype and mutant templates (five-to-six on average) and, therefore, we were unable to test the effect of the mutation alone without accounting for the mismatches. The impact of the mismatches alone revealed >2.1e7-fold loss in product amplification in repeat 1, compounded by an additional 2000-fold loss due to the mutation (SNP 52; Supplementary Fig. 9c). In repeats 2 and 3, the mismatches resulted in a >4.5e4- and 1.6e4-fold loss, respectively, and the presence of SNP caused an additional 4.6e6 (Tt_SNP_2) and 1.3e6 (Tt_SNP_3) fold-loss, respectively.

Considering the high prevalence of mismatches between the published primers and the genome assemblies, we sought to identify if this was specific to the genomes used or if the primers and the repeats they were designed from were broadly different. Detection of the canonical repeats (without any primer mismatches or SNPs) in the other two available genomes for *A. lumbricoides* (A_lumbricoides_Ecuador_v1_5_4; WormBase ParaSite [WBPS]) and *T. trichiura* (TTRE3.0; WBPS) revealed no *Ascaris* repeats with 100% similarity (of 26 repeats with 90-100% similarity) and in *T. trichiura,* only three matches with 100% similarity were detected (of 302 repeats with 90-100% similarity). Considering these assays were initially designed based on a consensus sequence[19,20,41], it is likely that these mismatches reflect false variants; given that these false variants are found in the primers themselves, these assays likely perform sub-optimally due to the priming of mismatched targets.

These results demonstrate that the occurrence of variants, which we have identified as variable in frequency amongst globally distributed STH populations, in qPCR targets used as diagnostics can significantly impact the detection of STHs. In practical terms, these variants caused a 3-to-10 cycle change in qPCR, resulting in many samples, particularly samples with low-intensity infections, yielding false negative results for STH.

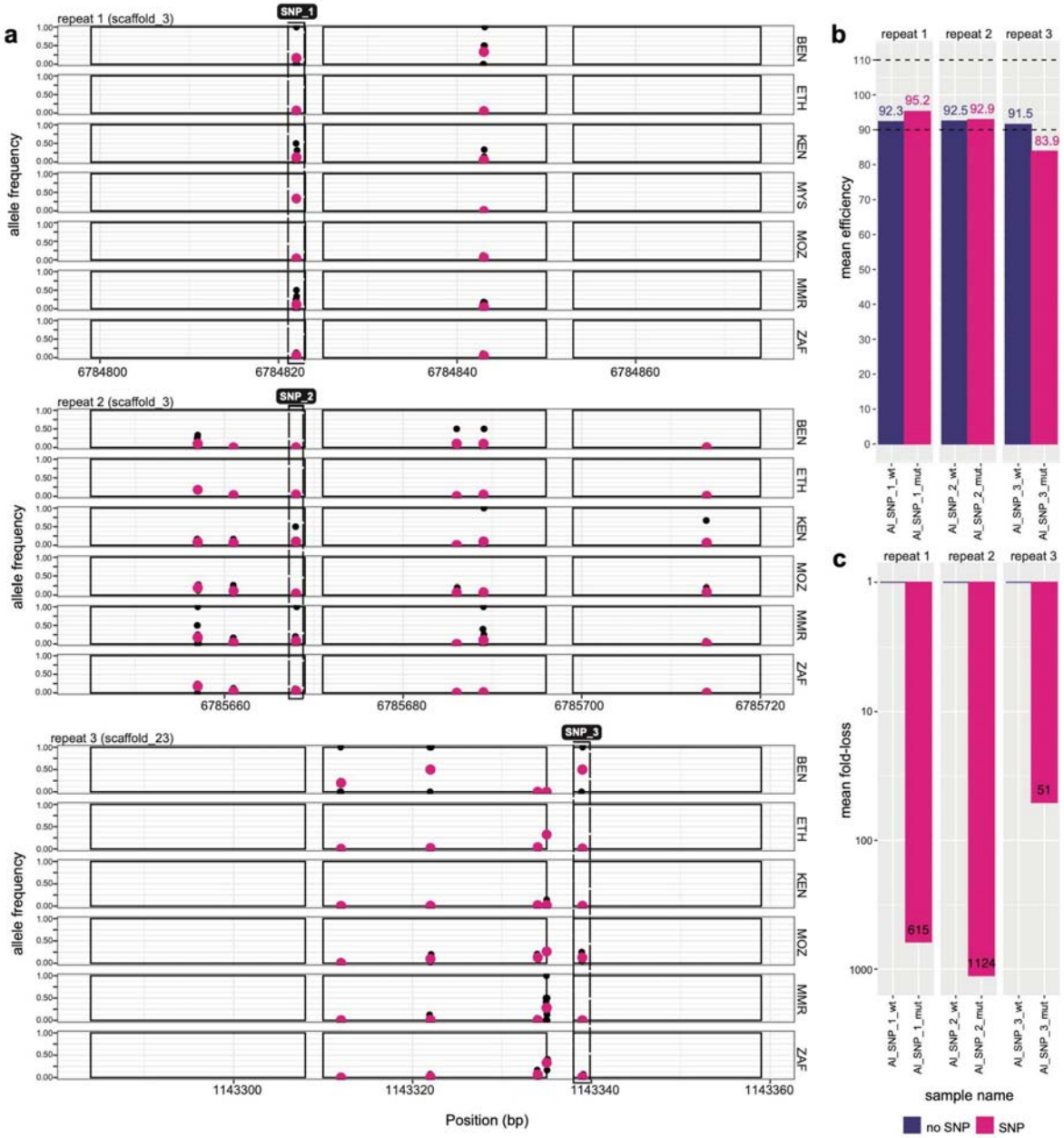


Fig. 4: Presence, distribution, and impact of genetic variation within diagnostic qPCR targets of *Ascaris lumbricoides*.

a, Shown are the genomic coordinates of three repeats - repeat 1, 2, and 3 - highlighting primer and probe binding sites (solid rectangles) used in the qPCR diagnostic test, the position of genetic variants (x-axis) - either individual samples (black points) or mean across samples (pink points) - and their frequency within each country (y-axis). Putative qPCR-disruptive variants found at the 3' end of the primer binding sites are depicted by dashed rectangles. **b,** qPCR efficiencies were determined by generating standard curves of five serial dilutions (100 pg/µl to 10 fg/µl) on each of the repeats in the absence ("wt") or

presence of the SNP ("mut"). The 90-110% dashed lines show the acceptable qPCR efficiency cutoffs. **c,** The mean fold-loss was calculated to assess the effect of the SNP in qPCR quantitation and product loss due to late amplification. Due to the large effects of the SNPs and other mismatches within the primer binding site of the wildtype samples, the mean fold loss of the wildtype is relative to the canonical repeat, whereas the mean fold loss of the mutant is relative to the wildtype repeat, within each assay. The mean normalised C_q difference was estimated from all serial dilutions. Country codes are as follows: BEN = Benin; ETH = Ethiopia; KEN = Kenya; ZAF = South Africa; MYS = Malaysia; MMR = Myanmar; MOZ = Mozambique.

Discussion

Soil-transmitted helminths impact a large portion of the global population, perpetuating poverty and disease in some of the most neglected communities worldwide. Effective and sustainable control of STHs requires a robust understanding of how parasite populations change over time. Here, we have addressed three key aspects of genomic epidemiology relevant to STH control, characterising (i) complex samples containing one or more helminth species, (ii) genetic diversity and connectivity of parasite populations, and (iii) the impact of genetic variation between STH populations on the performance and efficiency of conventional molecular diagnostics such as qPCR.

Understanding parasite genetic diversity will provide vital information for STH epidemiology, from their natural transmission patterns within and between populations, through population decline in response to drug treatment (including drug resistance) to recrudescence after termination of treatment. Although an increasing number of studies have focused on helminth genetic diversity[24,31,36,37,42], very little is known about the genetic diversity of STHs throughout their geographical range. For the first time, we have identified contrasting patterns of genetic variation and connectivity between globally distributed STH species. We note that due to the mixed nature of the sample cohort, it was difficult to directly compare the pooled egg or faecal samples to the individual worms, limiting our ability to assess population dynamics at a higher granularity. We also acknowledge that low-coverage sequencing poses some limitations. While the approach was specific, the sensitivity was lower than that of a targeted approach[43], which likely explains the higher frequency of *Ascaris* infections due to the greater biomass of eggs and the lack of *N. americanus* data despite being high in prevalence. Moreover, DNA sequencing from faecal samples meant that most of the reads were not helminth-derived and, therefore,

discarded. Nonetheless, we were able to detect population genetic signatures of *T. trichiura* consistent with increasing genetic differentiation correlated with increasing geographic distance due to limited dispersal. We also gained insights into likely zoonotic infections between pig-infective *A. suum* and human-infective *A. lumbricoides*. Although all samples analysed here were obtained from humans, there was an overwhelming bias towards *A. suum*-derived genetic diversity, leading us to question the evolutionary background of the original reference sample. Considering that *N. americanus* and *T. trichiura* are also zoonotic[24,42,44–46], genetic analysis of both human and animal reservoirs (monkeys, baboons, macaques for *Trichuris* and chimpanzees, gorillas for *Necator*) will become increasingly important for precisely understanding the populations targeted for control. These analyses need to be supported by new high-throughput molecular tools to characterise parasite diversity, including more sensitive sampling of the nuclear rather than mitochondrial genomes, and robust, consistent sampling approaches to improve the representitiveness and, therefore, genetic and geographic resolution of STH population genomics.

Understanding genetic variation, from single nucleotide polymorphisms to copy number variation of repetitive sequences, depends on reference genome resources, which differ in contiguity and completeness for STH species[47]. Our nuclear genome analyses revealed the position, number, and sequence similarity within species of these repeats in the genomes of *A. lumbricoides*, *N. americanus* and *T. trichiura*. However, our interpretation of these regions will change as genome assemblies improve, as evidenced by lower variation in copy number and high representation in the number of repeats in the newer, PacBio-derived *T. trichiura* genome. The availability of new genomes based on long-read sequencing will help to resolve the complexity of difficult-to-assemble repetitive regions of the genome (e.g.[48]) and, therefore, be more robust in identifying genetic variants in diverse helminth populations. New long read assemblies should also include diverse isolates to capture structural and geographic-specific diversity that may be missed by short-read sequencing alone.

The development of molecular approaches to diagnose and monitor STH infections offers great potential for increased sensitivity, specificity and throughput over conventional microscopy approaches. As such, qPCR-based diagnostics targeting repetitive sequences have been used in large-scale epidemiological and clinical trials to confirm elimination or monitor programmatic progress[14–17,49]. Genetic variation in diagnostic targets may impact monitoring of programs. If not accounted for, there is a significant risk of undiagnosed false negative infections if qPCR were applied in program settings. We demonstrate that, under *in vitro* controlled conditions, single

nucleotide genetic variants can significantly influence the ability of qPCR to detect diagnostic targets. We acknowledge that our *in vitro* assays are extreme scenarios and assume that all targets in a genome would be identical and perform equally poorly in the presence of the variant. Our analyses of repeat diversity show this to be false in reality. However, detecting the canonical repeat sequences within the reference genomes was challenging, suggesting that these established assays may perform sub-optimally. Redesigning diagnostic targets while accounting for genome-wide population genomic data will improve the assay by avoiding known variants and may allow the selection of more conserved targets less sensitive to disruptive genetic variants. Moreover, validation of new assays using a cohort of geographically diverse samples that are representative of continent, region, and/or country would provide a greater understanding of the strengths and weaknesses of the assay and, potentially, greater confidence in their application in new populations.

Our study emphasises the importance of sampling a wider range of human and nonhuman hosts globally to improve our understanding of the genetic makeup and evolution of STHs over time and space. Further work to define the genetic connectivity of STH populations may enable the definition of transmission zones and networks, providing a more precise means to prioritise control efforts. Our analyses provide a clear rationale for more comprehensive evaluation and validation of repetitive sequences currently being used as targets of molecular diagnostics. Collectively, these data and subsequent studies will form the foundations for the genomic epidemiology of all STH and their sustainable control as a public health problem.

Material and methods

Origin of samples and datasets

The complete list of samples, country of origin, and accession numbers are listed in Supplementary Table 1 and are further described in the Supplementary Information. A total of 1000 samples were used in the study, of which 850 (19 countries) were publicly available short-read shotgun metagenomics data obtained from the European Nucleotide Archive (ENA) from worm and faecal/egg samples known to have STH infections confirmed by either microscopy or qPCR, and 150 (13 countries) were DNA samples that underwent low-coverage genome sequencing, which were selected from a larger cohort of samples based on positive infections from microscopy data.

DNA extraction and high-throughput sequencing

As a consequence of collating a selection of diverse samples from multiple countries, a total of six different extraction kits were used to isolate DNA from the 150 faecal or worm samples (Supplementary Table 7) that were sequenced as part of this study. The faecal samples and/or concentrated eggs from Argentina (n = 1), Bangladesh (n = 10), Myanmar (n = 32), Kenya (n = 7), and South Africa (n = 7) were processed using the FastDNA Spin Kit for Soil (MP Biomedicals) and a high-speed homogeniser with modifications, including an extraction control (Papaiakovou et al., 2019b). The sample cohort from Benin (n = 25), Malawi (n = 25) and India (n = 25) was extracted as described elsewhere with a 96-well plate shaking system and a KingFisher Flex purification system using MagMax Microbiome Ultra Nucleic acid isolation kit [15]. The faecal samples from Ethiopia (n = 1) and the samples from the Democratic Republic of Congo (n = 1) were processed using the QIAsymphony DSP Virus/Pathogen Kit (QIAGEN) with modifications, including the use of a high-speed homogeniser (Vlaminck et al., 2018). The samples from Sri Lanka (n = 1) were processed using the PowerSoil DNA Isolation kit (Mobio, QIAGEN), including a negative/blank extraction control. The samples from Nigeria (n = 11) were extracted with the ZymoBIOMICS kit for DNA extraction (Zymo Research). The worm samples from Fiji (n = 2) and Ethiopia (n = 2) were extracted using the Isolate II Genomic DNA extraction kit (Bioline, Meridian Bioscience).

Sequencing libraries were prepared from the DNA extractions using the Plant and Animal Whole Genome Sequencing pipeline (Novogene, Hong Kong). One hundred fifty samples were

sequenced on a single lane of an Illumina NovaSeq 6000 (Novogene, Hong Kong) using 2×150 bp paired-end chemistry, generating approximately ten gigabases (Gb) of sequencing data per sample. Due to the initial low data yield of 10 samples, the libraries were sequenced further on two additional sequencing lanes. Individual sample accessions are described in Supplementary Table 1.

Raw read processing and mapping to reference genomes

Reference fasta files representing the mitochondrial genomes and whole genome assemblies of helminths previously reported to be found in the faeces of human hosts were collated from WBPS (release WBSP17;[50]) or NCBI (Supplementary Table 2). An improved, recently published genome assembly of *T. trichiura* [24] and the most recent genome of *A. lumbricoides*[37] were used instead of the assembly versions available from WBPS. The reference sequences were combined to generate a metagenomic reference set for each of the mitochondrial or whole genome sequence sets. By doing so, metagenomic sequencing reads composed of one or multiple species could map competitively and specifically to the reference genome of the species from which they were derived, enabling the generic classification of any sample.

Reference fasta files were indexed using Burrow-Wheeler Aligner (BWA) before trimmed sequencing reads (using *Trimmomatic*, v.0.39-2;[51]) were mapped using BWA-MEM (v.0.7.17-r1188; [52]). Unmapped reads were filtered using *samtools view*, retaining reads >80 bp in length. Mapped reads were further filtered to remove hard clipped (based on CIGAR pattern) reads using *samclip* (v.0.4.0), after which unique alignments were retained from the sorted BAM files using *sambamba* (v.0.8.2;[53]). The remaining soft-clipped alignments were filtered further by removing any reads with 'S' (indicating a soft-clipped alignment) on the CIGAR column of the BAM files. Lastly, duplicated reads were marked and removed from the BAM files using *Picard MarkDuplicates* (v.picard-2.18.29-0).

The number of mapped reads per sample per species was analysed using *bedtools multicov* on filtered BAM files. To compare diverse samples, read count normalisation was performed by dividing the number of reads mapped by the total number of reads per sample and then dividing by the mitochondrial genome size in Mb. A minimum coverage of 5X was selected for downstream analysis. Co-infections across all species and samples were visualised with the *UpSetR* package in R[54].

Mitochondrial genome variant calling

Variant calling was performed with *bcftools mpileup* (v.1.16). Two VCF files were generated, one containing only variant sites (used for pooled data, downstream, filtered for both min and max number of alleles = 2) and a second VCF that uses both variant and invariant sites (used for genotype-based population analysis in the individual worm data, filtered only for max alleles = 2). The only-variants-VCF file was further filtered to include only SNPs within coding genes (*vcftools --vcf* VCF *--bed* GENE.bed*)*. A data-driven approach was employed to filter variants based on removing tails (lower 10%) of variant distributions on QUAL metrics. Variants were further filtered to ensure (i) a maximum of two alleles, (ii) unique variants, (iii) only SNPs to be retained (any insertions/deletions were removed), and (iv) a per sample missingness > 0.7.

Population structure analysis

Broad-scale genetic relatedness between samples and populations was explored by analysing mitochondrial genome allele frequencies using principal component analysis from the R package *pcaMethods* (v.pcaMethods_1.90.0;[55]). Specifically, Bayesian principal component analysis (BPCA) was used to account for missing values in the dataset. Due to the presence of pooled and individual samples included in the analysis, allele frequencies rather than genotypes were used. To ensure that any existing population clustering was not attributed to differences in coverage or missingness in the data, each BPCA plot was coloured by normalised coverage and shaped by missingness in the data (total missingness present <0.3). Genetic differentiation between samples was performed using *Grenedalf* (v.0.3.0;[56]) for the pooled (faecal or concentrated eggs) samples and *pixy* (v.1.2.7.beta1;[57]) for the samples originating from individual worms.

Mitochondrial genome-wide genetic diversity analyses

Samples were split as ‘pools’ and ‘individuals’ depending on whether they originated from faecal samples (including concentrated eggs) or individual worms, from which nucleotide diversity (π) per population and pairwise genetic divergence (D_{XY}) were determined.

In a pooled sequencing context, π is the average pairwise difference between all reads in a sample; higher π values indicate a rich pool of genetic variation within a population, whilst lower π values indicate lower genetic diversity, possibly due to population bottlenecks, founder effects, and strong selection pressures. D_{XY} reflects the absolute divergence in π between two

populations; higher D_{XY} values indicate greater genetic divergence, reflecting potential isolation or selection pressures shaping the populations, whereas lower D_{XY} values suggest more genetic similarity, possibly due to recent shared ancestry or ongoing gene flow.

For the pools, π per population and pairwise D_{XY} were determined using *Grenedalf* (*grenedalf diversity --filter-sample-min-count 2 --filter-sample-min-coverage 2 --window-type chromosomes --pool-sizes 1000 --vcf-path; grenedalf fst --write-pi-tables --filter-sample-min-count 2 --filter-sample-min-coverage 1 --filter-region-list --window-type chromosomes --pool-sizes 1000 --method unbiased-hudson*). For the individuals, *pixy* was used (*pixy --stats pi dxy --vcf --chromosomes --populations --window_size 20000 --bypass_invariant_check 'yes'*). For the calculation of π and D_{XY} for the individuals, an all-sites VCF containing both variant and non-variant sites was used, whereas, for the pooled data, the original BAM files were used to avoid intrinsic biases in the VCF file from the variant caller (originally designed to be used with diploid data).

For this study, a mean D_{XY} value per country comparison was calculated. Variant frequency was also calculated for both pooled ($freq_{allele}$) and individual ($freq_{SNP}$) samples, as the number of variant positions (AC=1) divided by the total number of positions as a metric to compare individuals to pools.

Characterisation of the location, structure, and similarity of diagnostic repeats

A list of all the PCR-based diagnostic targets assessed in this study can be found in Supplementary Table 4. The different variations of each target in each genome assembly were assessed with *nucmer* (*mummer* package, v.3.23;[58]), allowing for a minimum similarity of 80% and cluster size of 15. The results were filtered further for 90% coverage and 90% similarity, and the *fasta* sequences from all similar targets were obtained using *bedtools getfasta*, aligned using *ClustalOmega* (*msa* package, R) and were visualised using *Heatmap.2* in R. To visualise the location and coverage of each target, each genome was split into 10 kb windows (with *bedtools makewindows*), and the coverage was calculated using *bedtools coverage*; results were plotted in R.

The number of primer and probe binding sites in each genome was assessed with *seqkit locate* (v.2.7.0), allowing up to five mismatches between the oligo and target sequences. The intersection between repeats and oligo binding sites was assessed with *bedtools intersect* and the resulting bed file was used to filter the downstream variants.

Repeat coverage per country

Repeat coverage was determined using *bedtools multicov* (with a minimum overlap of 0.51) in merged-by-country BAM files. Values were normalised by the length of each repeat and the total number of reads. *BUSCO* (v.5.6.1;[59]) was used to obtain the coordinates of single-copy exons and assess coverage in the country-merged BAM files, and a mean-per-country value was calculated. To obtain 'repeat copies' per country, original repeat coverage was divided by the mean-per-country single copy exon coverage and was visualised in R.

A similar approach was followed for the mitochondrial genomes. Raw read counts (per country) were divided by the length of the mitochondrial genome, after which they were further normalised with single-copy exon coverage to obtain mitochondrial genome copies per country. Mitochondrial genome copies and nuclear repeat copies were directly compared at a country and life-stage level (worms vs eggs). Statistical analysis was performed in R (v.4.2.2), using the Kruskal-Wallis test to measure the difference between the nuclear and mitochondrial copies, with $p < 0.05$ being considered a significant difference.

Variant calling in repetitive sequences

Variant calling was performed with *bcftools mpileup* (v.1.16), generating a single VCF file per species for all the targets tested. Each VCF file was filtered for minimum and maximum number of alleles, and unique SNPs (indels were excluded), and only positions within-binding sites of primer and probe oligos were retained. SNPs were filtered for quality, sum depth for alternate and reference alleles and sample missingness (SNPs with values below the 10th quantile were removed for all the criteria above). Different degrees of per-site-missingness were also explored, and a threshold of 0.7 was selected. The number of SNPs surviving each filter (and the number of repeats) is shown in Supplementary Table 5.

SNP selection for validation

For SNP validation, we focused on variants in the 3' end (up to 5 bp) of the qPCR oligos with variable frequencies (and higher allele depth) in most countries. Three SNPs per species were selected for testing. To accurately assess the effect of the SNP on qPCR amplification of the repeat, we tested the repeats containing the identified SNPs along with the wild-type version of the repeat (without the SNP). To do so, we synthesised (i) the wild type (wt) repeat target (no SNP present) (ii) and the mutated (mut) repeat targets (with the SNP) in pIDTSMART-AMP

plasmids (Integrated DNA Technologies, UK). The complete sequences tested can be found in Supplementary Table 6. We also tested the published canonical repeats from which the original assays were designed for comparison.

Assessing the impact of variants using qPCR standard curves

To determine the effect of variants on the qPCR amplification efficiency, all targets were assessed by generating a 5-fold serial dilution standard curve, in triplicate, from 100 pg/ul to 10 fg/ul. qPCR assays were set up for all targets using TaqPath ProAmp Master Mix (Applied Biosystems) in a final volume of 7 µl, containing 5 µl of master mix (reagent mix, forward and reverse primers, probe and nuclease-free water) and 2 µl of template DNA. To amplify the *A. lumbricoides* targets, the forward primer (CTTGTACCACGATAAAGGGCAT; 125 nM final concentration), reverse primer (TCCCTTCCAATTGATCATCGAATAA; 500 nM), and probe (TCTGTGCATTATTGCTGCAATTGGGA; 125 nM) were used. To amplify the *T. trichiura* targets, the forward primer (GGCGTAGAGGAGCGATTT; 62.5 nM), reverse primer (TACTACCCATCACACATTAGCC; 250 nM), and probe (TTTGCGGGCGAGAACGGAAATATT; 125 nM) were used. The probes contained at 5'-FAM-label (double quenched ZEN/3IABkFQ chemistry; IDT Technologies). Cycling was performed on a ThermoFisher StepOnePlus instrument, using an initial 95°C for 10 min followed by 40 cycles of 95°C for 15 sec for denaturation and 59°C for 1 min for annealing and extension.

From the standard curve data, amplification efficiency values between 90-110% were considered to be within an acceptable range and to which outlier values of higher or lower efficiency were measured. Mean fold-change was measured as $2^{\text{abs}((\text{wild_type_Cq} * \text{wild_type_efficiency}) - (\text{mutated_Cq} * \text{mutated_efficiency}))}$ to assess the effect of the SNP in qPCR quantitation and product loss due to late amplification.

Data availability

Supplementary Table 1 describes the study and sample accessions for the sequencing data generated and analysed in this study, which is available from ENA. Reference genomes used for read mapping and comparative genomics are described in Supplementary Table 2.

Code availability

The code used in the study is available via GitHub at https://github.com/MarinaSci/Global_skim_analysis.

References

1. Bethony, J. *et al.* Soil-transmitted helminth infections: ascariasis, trichuriasis, and hookworm. *Lancet* **367**, 1521–1532 (2006).
2. Else, K. J. *et al.* Whipworm and roundworm infections. *Nat Rev Dis Primers* **6**, 44 (2020).
3. Gordon, C. A. *et al.* Strongyloidiasis. *Nat Rev Dis Primers* **10**, 6 (2024).
4. Campbell, S. J. *et al.* Complexities and Perplexities: A Critical Appraisal of the Evidence for Soil-Transmitted Helminth Infection-Related Morbidity. *PLoS Negl. Trop. Dis.* **10**, e0004566 (2016).
5. Yap, P., Utzinger, J., Hattendorf, J. & Steinmann, P. Influence of nutrition on infection and re-infection with soil-transmitted helminths: a systematic review. *Parasit. Vectors* **7**, 229 (2014).
6. Montresor, A. *et al.* The global progress of soil-transmitted helminthiases control in 2020 and World Health Organization targets for 2030. *PLoS Negl. Trop. Dis.* **14**, e0008505 (2020).
7. Vaz Nery, S. *et al.* The role of water, sanitation and hygiene interventions in reducing soil-transmitted helminths: interpreting the evidence and identifying next steps. *Parasit. Vectors* **12**, 273 (2019).
8. WHO | Ending the neglect to attain the Sustainable Development Goals: A road map for neglected tropical diseases 2021–2030. (2020).
9. Stuyver, L. J. & Levecke, B. The role of diagnostic technologies to measure progress toward WHO 2030 targets for soil-transmitted helminth control programs. *PLoS Negl. Trop. Dis.* **15**, e0009422 (2021).
10. Katz, N., Chaves, A. & Pellegrino, J. A simple device for quantitative stool thick-smear technique in *Schistosomiasis mansoni*. *Rev. Inst. Med. Trop. Sao Paulo* **14**, 397–400 (1972).

11. Nikolay, B., Brooker, S. J. & Pullan, R. L. Sensitivity of diagnostic tests for human soil-transmitted helminth infections: a meta-analysis in the absence of a true gold standard. *Int. J. Parasitol.* **44**, 765–774 (2014).
12. Barros, N. & Montes, M. Infection and Hyperinfection with Strongyloides stercoralis: Clinical Presentation, Etiology of Disease, and Treatment Options. *Current Tropical Medicine Reports* **1**, 223–228 (2014).
13. Papaiakovou, M., Littlewood, D. T. J., Gasser, R. B. & Anderson, R. M. How qPCR complements the WHO roadmap (2021-2030) for soil-transmitted helminths. *Trends Parasitol.* **37**, 698–708 (2021).
14. Benjamin-Chung, J. *et al.* Comparison of multi-parallel qPCR and double-slide Kato-Katz for detection of soil-transmitted helminth infection among children in rural Bangladesh. *PLoS Negl. Trop. Dis.* **14**, e0008087 (2020).
15. Pilotte, N. *et al.* Development and validation of a high-throughput qPCR platform for the detection of soil-transmitted helminth infections. *bioRxiv* (2023) doi:10.1101/2023.11.27.23299079.
16. Hasegawa, M. *et al.* What does soil-transmitted helminth elimination look like? Results from a targeted molecular detection survey in Japan. *Parasit. Vectors* **13**, 6 (2020).
17. Pickering, A. J. *et al.* Effects of single and integrated water, sanitation, handwashing, and nutrition interventions on child soil-transmitted helminth and Giardia infections: A cluster-randomized controlled trial in rural Kenya. *PLoS Med.* **16**, e1002841 (2019).
18. Easton, A. V. *et al.* Multi-parallel qPCR provides increased sensitivity and diagnostic breadth for gastrointestinal parasites of humans: field-based inferences on the impact of mass deworming. *Parasit. Vectors* **9**, 38 (2016).
19. Pilotte, N. *et al.* Targeting a highly repeated germline DNA sequence for improved real-time PCR-based detection of *Ascaris* infection in human stool. *PLoS Negl. Trop. Dis.* **13**, e0007593 (2019).

20. Pilotte, N. *et al.* Improved PCR-Based Detection of Soil Transmitted Helminth Infections Using a Next-Generation Sequencing Approach to Assay Design. *PLoS Negl. Trop. Dis.* **10**, e0004578 (2016).
21. Papaiakovou, M. *et al.* A novel, species-specific, real-time PCR assay for the detection of the emerging zoonotic parasite *Ancylostoma ceylanicum* in human stool. *PLoS Negl. Trop. Dis.* **11**, e0005734 (2017).
22. Mejia, R. *et al.* A novel, multi-parallel, real-time polymerase chain reaction approach for eight gastrointestinal parasites provides improved diagnostic capabilities to resource-limited at-risk populations. *Am. J. Trop. Med. Hyg.* **88**, 1041–1047 (2013).
23. Hii, S. F. *et al.* Development and Evaluation of a Multiplex Quantitative Real-Time Polymerase Chain Reaction for Hookworm Species in Human Stool. *Am. J. Trop. Med. Hyg.* **99**, 1186–1193 (2018).
24. Doyle, S. R. *et al.* Population genomics of ancient and modern *Trichuris trichiura*. *Nat. Commun.* **13**, 1–12 (2022).
25. Choi, Y.-J. *et al.* Genomic diversity in *Onchocerca volvulus* and its *Wolbachia* endosymbiont. *Nat Microbiol* **2**, 16207 (2016).
26. Small, S. T. *et al.* Human Migration and the Spread of the Nematode Parasite *Wuchereria bancrofti*. *Mol. Biol. Evol.* **36**, 1931–1941 (2019).
27. Durrant, C. *et al.* Population genomic evidence that human and animal infections in Africa come from the same populations of *Dracunculus medinensis*. *PLoS Negl. Trop. Dis.* **14**, e0008623 (2020).
28. Shortt, J. A. *et al.* Population genomic analyses of schistosome parasites highlight critical challenges facing endgame elimination efforts. *Sci. Rep.* **11**, 6884 (2021).
29. Platt, R. N., 2nd *et al.* Genomic analysis of a parasite invasion: Colonization of the Americas by the blood fluke *Schistosoma mansoni*. *Mol. Ecol.* **31**, 2242–2263 (2022).
30. Sallé, G. *et al.* The global diversity of *Haemonchus contortus* is shaped by human

intervention and climate. *Nat. Commun.* **10**, 4811 (2019).

31. Gandasegui, J. *et al.* Genome structure and population genomics of the canine heartworm *Dirofilaria immitis*. *Int. J. Parasitol.* (2023) doi:10.1016/j.ijpara.2023.07.006.
32. Han, L. *et al.* Chromosome-scale assembly and whole-genome sequencing of 266 giant panda roundworms provide insights into their evolution, adaptation and potential drug targets. *Mol. Ecol. Resour.* **22**, 768–785 (2022).
33. Thorn, C. S., Maness, R. W., Hulke, J. M., Delmore, K. E. & Criscione, C. D. Population genomics of helminth parasites. *J. Helminthol.* **97**, e29 (2023).
34. Crellen, T. *et al.* Whole genome resequencing of the human parasite Schistosoma mansoni reveals population history and effects of selection. *Sci. Rep.* **6**, 1–13 (2016).
35. Park, Y. C., Kim, W. & Park, J.-K. The complete mitochondrial genome of human parasitic roundworm, *Ascaris lumbricoides*. *Mitochondrial DNA* **22**, 91–93 (2011).
36. Betson, M., Nejsum, P., Bendall, R. P., Deb, R. M. & Stothard, J. R. Molecular epidemiology of ascariasis: a global perspective on the transmission dynamics of *Ascaris* in people and pigs. *J. Infect. Dis.* **210**, 932–941 (2014).
37. Easton, A. *et al.* Molecular evidence of hybridization between pig and human *Ascaris* indicates an interbred species complex infecting humans. *Elife* **9**, (2020).
38. Okimoto, R., Macfarlane, J. L., Clary, D. O. & Wolstenholme, D. R. The mitochondrial genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum*. *Genetics* **130**, 471–498 (1992).
39. Estrem, B. & Wang, J. Programmed DNA elimination in the parasitic nematode Ascaris. *PLoS Pathog.* **19**, e1011087 (2023).
40. Grant, J. R., Pilotte, N. & Williams, S. A. A Case for Using Genomics and a Bioinformatics Pipeline to Develop Sensitive and Species-Specific PCR-Based Diagnostics for Soil-Transmitted Helminths. *Front. Genet.* **10**, 883 (2019).
41. Novák, P., Neumann, P. & Macas, J. Global analysis of repetitive DNA from unassembled

sequence reads using RepeatExplorer2. *Nat. Protoc.* **15**, 3745–3776 (2020).

42. Rivero, J., Cutillas, C. & Callejón, R. *Trichuris trichiura* (Linnaeus, 1771) From Human and Non-human Primates: Morphology, Biometry, Host Specificity, Molecular Characterization, and Phylogeny. *Front Vet Sci* **7**, 626120 (2020).
43. Papaiakovou, M. *et al.* Evaluation of genome skimming to detect and characterise human and livestock helminths. *Int. J. Parasitol.* **53**, 69–79 (2023).
44. Ilík, V. *et al.* High diversity and sharing of strongylid nematodes in humans and great apes co-habiting an unprotected area in Cameroon. *PLoS Negl. Trop. Dis.* **17**, e0011499 (2023).
45. Hasegawa, H. *et al.* Humans and great apes cohabiting the forest ecosystem in central african republic harbour the same hookworms. *PLoS Negl. Trop. Dis.* **8**, e2715 (2014).
46. Hawash, M. B. F., Andersen, L. O., Gasser, R. B., Stensvold, C. R. & Nejsum, P. Mitochondrial Genome Analyses Suggest Multiple *Trichuris* Species in Humans, Baboons, and Pigs from Different Geographical Regions. *PLoS Negl. Trop. Dis.* **9**, e0004059 (2015).
47. Doyle, S. R. Improving helminth genome resources in the post-genomic era. *Trends Parasitol.* **38**, 831–840 (2022).
48. Brann, T. *et al.* Subtelomeric plasticity contributes to gene family expansion in the human parasitic flatworm *Schistosoma mansoni*. *BMC Genomics* **25**, 217 (2024).
49. Dunn, J. C. *et al.* The increased sensitivity of qPCR in comparison to Kato-Katz is required for the accurate assessment of the prevalence of soil-transmitted helminth infection in settings that have received multiple rounds of mass drug administration. *Parasit. Vectors* **13**, 324 (2020).
50. Howe, K. L., Bolt, B. J., Shafie, M., Kersey, P. & Berriman, M. WormBase ParaSite - a comprehensive resource for helminth genomics. *Mol. Biochem. Parasitol.* **215**, 2–10 (2017).
51. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
52. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

arXiv [q-bio.GN] (2013).

53. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
54. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
55. Stacklies, W., Redestig, H., Scholz, M., Walther, D. & Selbig, J. pcaMethods--a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* **23**, 1164–1167 (2007).
56. Czech, L., Spence, J. P. & Expósito-Alonso, M. grenedalf: population genetic statistics for the next generation of pool sequencing. *arXiv [q-bio.PE]* (2023).
57. Korunes, K. L. & Samuk, K. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol. Ecol. Resour.* **21**, 1359–1368 (2021).
58. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
59. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).

Acknowledgements

The authors would like to thank John M Colford Jr for the support providing the samples from Bangladesh, Marisa Juarez and Paola A. Vargas for their help with egg collection and logistical support during the field sample processing in Orán, Argentina, the NHM Molecular Labs, especially Ranbir Bailey, Elena Lugli and Claire Griffin, for access and logistical support, Stefano Colombo for all of their help with DNA extractions on the samples from Uganda, and the Helminth Genomics group at the Wellcome Sanger Institute for insightful discussions throughout this project.

Sequencing of samples was supported by the Bill & Melinda Gates Foundation (USA; OPP1129535) and the Wellcome Trust (UK; 206194). MP is supported by a Harding Distinguished Postgraduate Scholarship held at the University of Cambridge (UK). JBC is a Chan Zuckerberg Biohub Investigator. SRD is supported by a UKRI Future Leaders Fellowship (UK; MR/T020733/1) and the Wellcome Trust (UK; 206194). Samples provided by Ghent University (Belgium) was part of field trials supported by the Bill & Melinda Gates Foundation (USA; OPP1120972). Samples from Kenya were collected with financial support from the University of Georgia Research Foundation, Inc., which was funded by the Bill & Melinda Gates Foundation (USA; INV-010428) for the SCORE Project. Samples from Mozambique were collected by the Centro de Investigação em Saúde de Manhiça (CISM), which is supported by the Government of Mozambique and the Spanish Agency for International Development Cooperation (AECID) and Institut de Salut Global de Barcelona (ISGlobal), which receives support from the grant CEX2018-000806-S funded by MCIN/AEI/10.13039/501100011033, and support from the Generalitat de Catalunya through the CERCA Program. For the purpose of Open Access, the authors have applied a CC BY public copyright Licence to any Author Accepted Manuscript version arising from this submission.

Author information

These authors contributed equally: D. Timothy J. Littlewood and Cinzia Cantacessi.

Contributions

MP, DTJL and SRD conceived and designed the study; SRD, DTJL, and CC co-supervised the study; SRA, OA, RMA, RB, JBC, MCP, NRC, DC, ROC, PC, AC, JD, SG, JG , BGP, EH, MI, TPJ, KK, EFK, AJK, BL, AJL, AM, IM, MM, MMV, ZM, AM Jr, RM, HM, OM , JM, PM, VN , MRO, CS, JLW, SWM, SAW provided materials; MP extracted DNA, prepared samples for sequencing, and performed qPCR analyses; MCP and JG prepared pooled samples from Mozambique and extracted DNA; MP led and performed the bioinformatics analyses; MP and SRD analysed and interpreted the results, with input from contributions from AW, CC, and DTJL; MP and SRD drafted the original paper, with contributions from AW, CC, and DTJL. All authors contributed to the revision and approval of the final paper.

Corresponding authors

Correspondence to Marina Papaiakovou and Stephen R. Doyle.

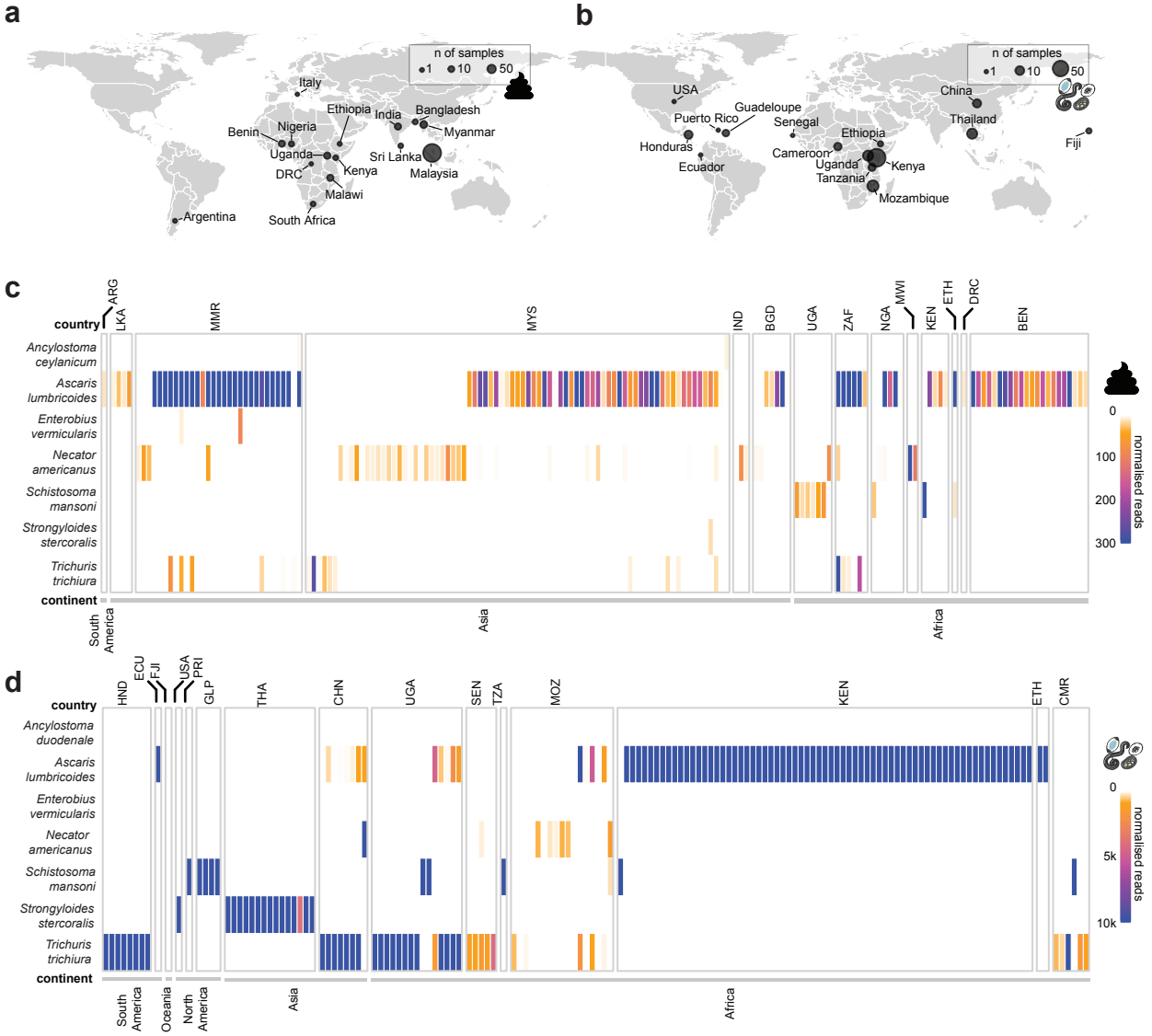
Ethics declarations

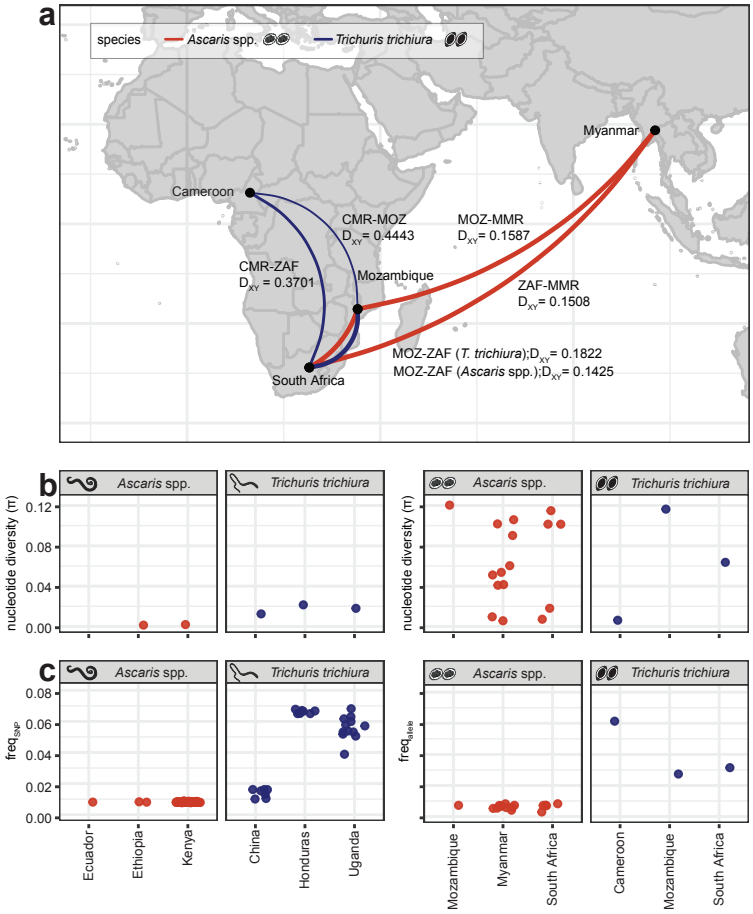
Competing interests

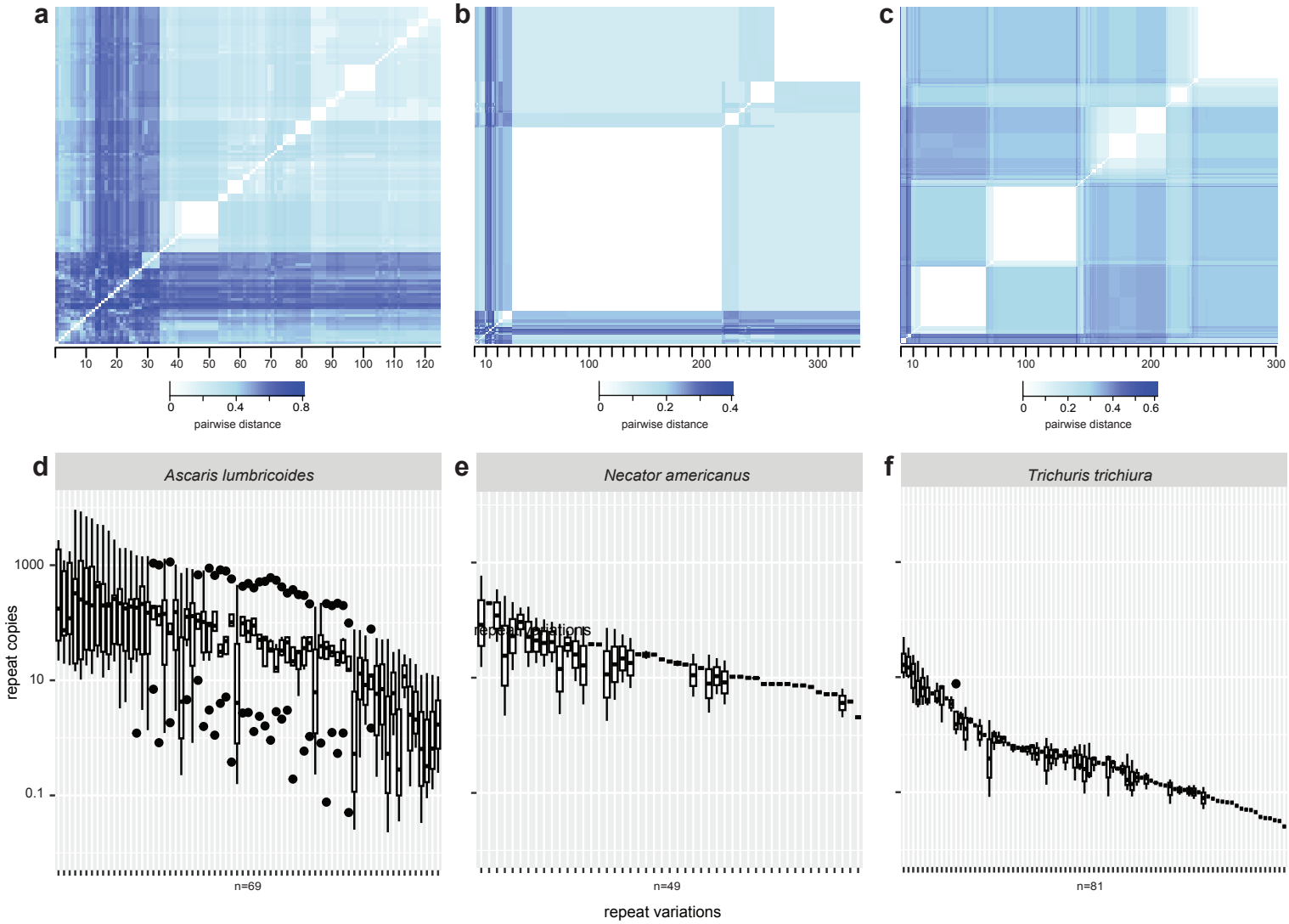
The authors declare no competing interests.

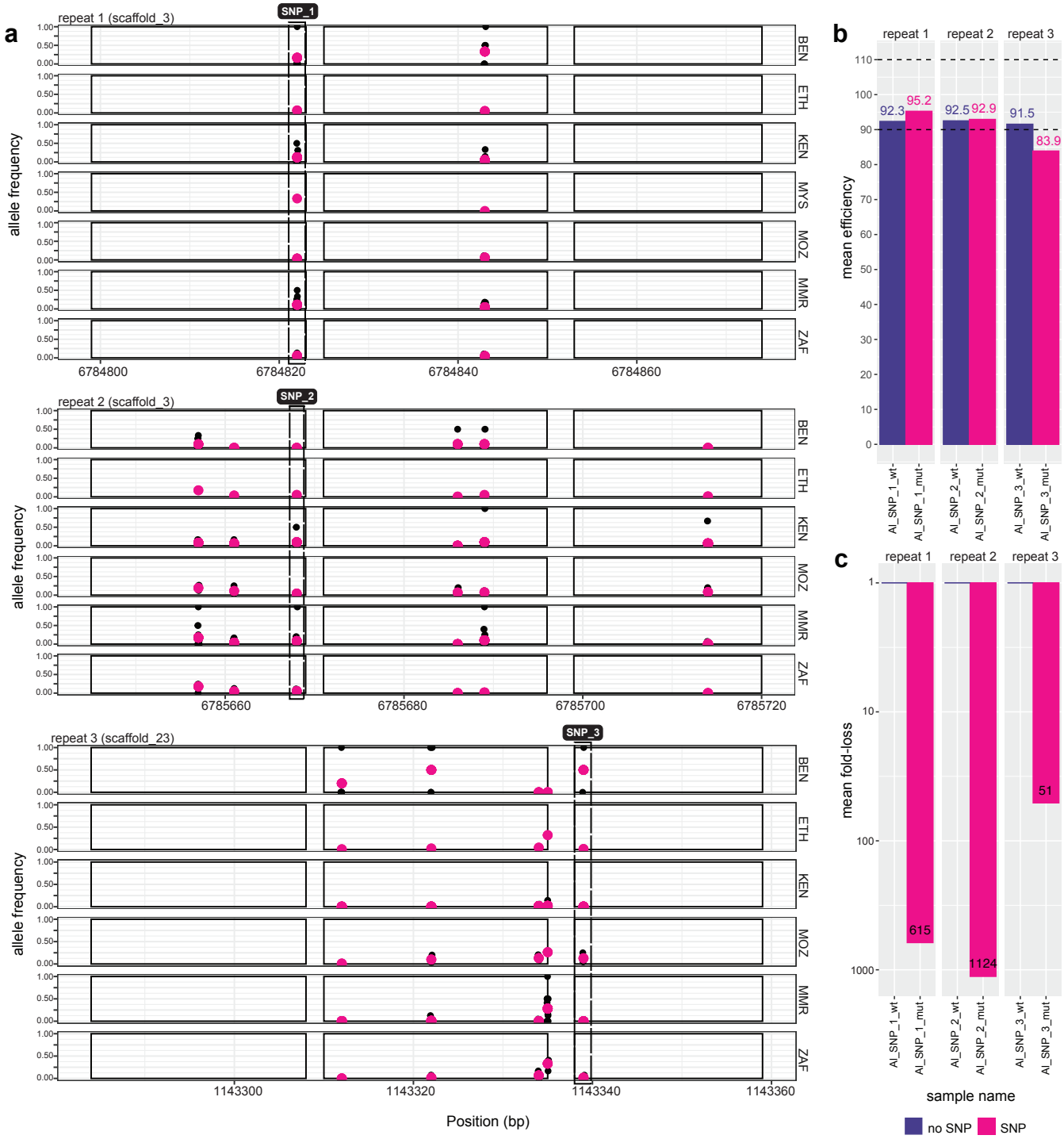
Supplementary Information

- Supplementary Figs. 1-9
- Supplementary Tables 1-7


a
n of samples
1
10
50
Italy
Ethiopia
India
Bangladesh
Myanmar
Benin
Nigeria
Uganda
Sri Lanka
DRC
Kenya
Malaysia
Malawi
Argentina
South Africa
b
n of samples
1
10
50
USA
Guadeloupe
Puerto Rico
Senegal
Honduras
Ecuador
Ethiopia
Cameroon
Uganda
Tanzania
Kenya
Mozambique
China
Thailand
Fiji
c
country
ARG
LKA
MMR
MYS
IND
BGD
UGA
ZAF
NGA
MWI
KEN
ETH
DRC
BEN
Ancylostoma ceylanicum
Ascaris lumbricoides
Enterobius vermicularis
Necator americanus
Schistosoma mansoni
Strongyloides stercoralis
Trichuris trichiura
continent
South America
Asia
Africa
0
100
200
300
normalised reads
d
country
HND
ECU
FJI
USA
PRI
GLP
THA
CHN
UGA
SEN
TZA
MOZ
KEN
ETH
CMR
Ancylostoma duodenale
Ascaris lumbricoides
Enterobius vermicularis
Necator americanus
Schistosoma mansoni
Strongyloides stercoralis
Trichuris trichiura
continent
South America
Oceania
North America
Asia
Africa
0
5k
10k
normalised reads


a
species
Ascaris spp.
Trichuris trichiura
Myanmar
Cameroon
CMR-MOZ
D_XY = 0.4443
MOZ-MMR
D_XY = 0.1587
CMR-ZAF
D_XY = 0.3701
Mozambique
ZAF-MMR
D_XY = 0.1508
MOZ-ZAF (T. trichiura);D_XY = 0.1822
MOZ-ZAF (Ascaris spp.);D_XY = 0.1425
South Africa
b
Ascaris spp.
Trichuris trichiura
nucleotide diversity (π)
0.12
0.08
0.04
0.00
c
freq_SNP
freq_alele
0.08
0.06
0.04
0.02
0.00
Ecuador
Ethiopia
Kenya
China
Honduras
Uganda
Mozambique
Myanmar
South Africa
Cameroon
Mozambique
South Africa


a
b
c
10 20 30 40 50 60 70 80 90 100 110 120
0 0.4 0.8
pairwise distance
10 100 200 300
0 0.2 0.4
pairwise distance
10 100 200 300
0 0.2 0.4 0.6
pairwise distance
d
Ascaris lumbricoides
e
Necator americanus
f
Trichuris trichiura
repeat copies
1000
10
0.1
n=69
n=49
n=81
repeat variations


a
repeat 1 (scaffold_3)
SNP_1
allele frequency
BEN
ETH
KEN
MYS
MOZ
MMR
ZAF
6784800
6784820
6784840
6784860
repeat 2 (scaffold_3)
SNP_2
6785660
6785680
6785700
6785720
repeat 3 (scaffold_23)
SNP_3
1143300
1143320
1143340
1143360
Position (bp)
b
repeat 1
repeat 2
repeat 3
mean efficiency
92.3
95.2
92.5
92.9
91.5
83.9
AI_SNP_1_wt
AI_SNP_1_mut
AI_SNP_2_wt
AI_SNP_2_mut
AI_SNP_3_wt
AI_SNP_3_mut
c
mean fold-loss
615
1124
51
sample name
no SNP
SNP