bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Extensive hybridization between pig and human *Ascaris* identifies a highly interbred species complex infecting humans

Alice V. Easton[1,2]*

Shenghan Gao[3,4]*

Scott P Lawton[5]

Sasisekhar Bennuru[1]

Asis Khan[6]

Eric Dahlstrom[7]

Rita G Oliveira[2]

Stella Kepha[8]

Steve F Porcella[7]

Joanne P Webster[2,9]

Roy M Anderson[2]

Michael E. Grigg[6]

Richard E Davis[3]**

Jianbin Wang[3,10]**

Thomas B Nutman[1]**

*Contributed equally to the manuscript

**Co-corresponding authors

richard.davis@cuanschutz.edu

jianbin.wang@utk.edu

tnutman@niaid.nih.gov

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Helminth Immunology Section, Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, Maryland 20892, USA.
2. Department of Infectious Disease Epidemiology, Imperial College London, London W2 1PG, UK.
3. Department of Biochemistry and Molecular Genetics, RNA Bioscience Initiative, University of Colorado School of Medicine, Aurora, Colorado 80045, USA.
4. Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.
5. Molecular Parasitology Laboratory, School of Life Sciences, Pharmacy & Chemistry, Kingston University London, Kingston Upon Thames, Surrey KT1 2EE, UK.
6. Molecular Parasitology Section, Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, Maryland 20892, USA
7. Genomics Unit, Research Technologies Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT 59840, USA.
8. London School of Tropical Medicine and Hygiene, Keppel St, Bloomsbury, London, WCIE 7HT, UK.
9. Royal Veterinary College, University of London, Department of Pathobiology and Population Sciences, Hertfordshire AL9 7TA, UK.
10. Present Address: Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, Tennessee 37996, USA

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Human ascariasis is a major neglected tropical disease caused by the nematode *Ascaris lumbricoides*. We report a 296 megabase (Mb) reference quality genome comprised of 17902 protein-coding genes derived from a single, representative *Ascaris* worm collected from 60 human hosts in Kenyan villages where pig husbandry is rare. Notably, the majority of human isolates (63/68) possessed mitochondrial genomes that clustered closer to the pig parasite *Ascaris suum* than to *A. lumbricoides*. Comparative phylogenomic analyses identified over 11 million nuclear-encoded SNPs but just two distinct genetic types that had recombined across the genomes analysed. The nuclear genomes had extensive heterozygosity and all samples existed as genetic mosaics with either *A. suum*-like or *A. lumbricoides*-like inheritance patterns supporting a highly interbred *Ascaris* species genetic complex. As no barriers appear to exist for anthroponotic transmission of these "hybrid" worms, a one-health approach to control the spread of human ascariasis will be necessary.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Approximately 447 million people were estimated to be infected with the intestinal nematode *Ascaris lumbricoides* in 2017, resulting in an estimated 3206 deaths and a loss of over 860,000 Disability-Adjusted Life Years (DALYs, Global Burden of Disease Study 2017 http://ghdx.healthdata.org/gbd-2017). Many infections go undiagnosed, but like other soil-transmitted helminths (STH), *Ascaris* spp. infections contribute significantly to global DALYs, perpetuating the cycle of poverty in areas of endemic infection [1–4]. Despite the large global burden of STH, little is known about *A. lumbricoides* transmission patterns or the true prevalence of infection with the pig parasite *A. suum* infection in people in endemic regions.

Deworming has become more widespread in areas of endemic STH infection [5]. Regional health authorities and global health organizations are now looking for strategies to build on these programs by achieving local elimination of STH as a public health problem [6]. A greater understanding of transmission dynamics (including the frequency of zoonotic transmission) using molecular epidemiological methods in settings where *A. lumbricoides* prevalence is low but persistent could help move current efforts towards successfully eliminating transmission through more targeted treatment.

Population genetic studies of *A. lumbricoides* have drawn varying conclusions about whether zoonotic transmission is frequent [7–10]. Some studies have shown that cross-species transmission occurs between pigs and humans living in close proximity [9,11–16]. This is especially common in non-endemic regions, probably because zoonotic transmission is less likely to be identified in areas where human-to-human transmission is common. To date, published results fail to conclude whether the human parasite *A. lumbricoides* and the pig parasite *Ascaris suum* are capable of interbreeding, and it is generally accepted that they exist as separate species. Furthermore, it is unclear whether pigs are an important reservoir of infection in humans worldwide or if *A. suum* is readily transmitted anthroponotically [7,17–20]. Studies have generally concluded that the genetic differences between *Ascaris* worms collected from human populations in different parts

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of the world [11,21] are the result of geographic reproductive isolation. Previous studies using *Ascaris* mitochondrial genomes or genes suggest there are *A. lumbricoides*-type (human-associated) and *A. suum*-type (pig-associated) clades [8,22,23]. Other work suggests multiple clades of worms, only one of which is unique to pigs [24].

In the current study, we constructed a reference-quality *Ascaris* genome (ALV5) based on sequences from a single female worm collected from a single person in Kenya. This person was presumed to be infected with *A. lumbricoides* as there is a lack of local pig husbandry. Draft *A. suum* genomes have previously been constructed from worms obtained from pigs in Australia [25] and in the United States [26,27]. The *Ascaris* genome ALV5 was found to be highly similar (99% identity) to the *A. suum* genome from worms collected from pigs in the United States [28]. Our mitochondrial and whole-genome analyses from an additional 68 individual worms indicate that *A. suum* and *A. lumbricoides* form a genetic complex that is capable of interbreeding. Our data support a model for a recent worldwide, multi-species *Ascaris* population expansion caused by the movement of humans and/or livestock globally. *Ascaris* isolates from both pigs and humans may be important in human disease, necessitating a one-health approach to control the spread of human ascariasis.

Results:

Human *Ascaris* reference genome to promote comparative genomic analyses

To generate a human *Ascaris* spp. germline genome assembly (prior to programmed DNA elimination [28]), ovarian DNA was sequenced from a single female worm collected from a Kenyan who was presumed to be infected with *A. lumbricoides* using Illumina paired-end and mate-pair libraries of various insert sizes with a total sequence coverage of ~27-fold (Table S1). Using these data, three different assembly strategies were used (see below and supplemental information).

The *de novo* assembly and semi-*de novo* strategies produced poor *A. lumbricoides* germline draft genomes (Table 1 and supplementary text). In the semi-*de novo* assembly, the majority of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

>4000 short contigs (making up 15.4 Mb of sequence) that could not be incorporated into the semi-*de novo* assembly are sequences that aligned to the genome at multiple positions. Comparison of the *A. suum* gene annotations to this assembly revealed a low *A. lumbricoides* gene number and high numbers of partial and split genes (Table 1, see footnote 3). These characteristics are typical of highly fragmented genomes or genomes with high levels of mis-assemblies [28].

Mapping of the human *Ascaris* reads to the *A. suum* reference genome [28] revealed an exceptionally high sequence similarity (>99% identity) between the two species with few human *Ascaris* reads that could not be mapped to *A. suum*. Based on this high sequence similarity, a third reference-based-only assembly strategy was used to generate the human *Ascaris* germline genome assembly using the *A. suum* germline genome as a reference (see methods). This approach led to a reference-quality human *Ascaris* genome assembly with many fewer gaps (only 0.98 Mb of sequence) and no unplaced contigs. The *Ascaris* genome assembled into 415 scaffolds with a combined size of 296 Mb. An additional 15.4 Mb of sequence was present in 4072 unscaffolded short contigs. The assembly N50 value was 4.63 Mb, with the largest scaffold measuring 13.2 Mb. The largest 50 scaffolds combined to represent 78% of the genome. The assembly was further polished using additional Illumina reads from the same worm to more accurately reflect single base differences, indels, and any potential local mis-assembled regions.

To evaluate the quality of the assembled genome, we mapped the *Ascaris* Illumina reads back to the reference-based *Ascaris* genome assembly and found that > 99% of the Illumina reads could be mapped, indicating that the reference-based assembly excluded very few *Ascaris* reads. We then mapped and transferred the extensive set of *A. suum* transcripts [25,28] to the human *Ascaris* germline assembly to annotate the genome, identifying and classifying 17902 protein-coding genes (Table 1, Table S2). As this reference-based assembly exhibits the best assembly attributes, including high continuity with a large N50, low gaps and unplaced sequences, and high-quality protein-coding genes (see Table 1), we suggest that this version should be used as a reference germline genome for a human *Ascaris* spp. isolate (available in NCBI GenBank with accession

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

number PRJNA515325). The other two assemblies are available and are discussed in more detail in the supplemental text.

Like *A. suum* embryos, *A. lumbricoides* embryos undergo programmed DNA elimination during the differentiation of the somatic cells from the germline in early development [29,30]. In *A. suum*, ~30 Mb of 120 bp tandem repeats and ~1000 germline-expressed genes are lost from the germline to form the somatic genome [27,28]. We also sequenced the somatic genome from the intestine of the same female *A. lumbricoides* worm. Comparison of the germline and somatic genomes revealed that DNA elimination in the human *Ascaris* isolate (including the breaks, sequences, and genes eliminated) was identical to that described for the pig *A. suum* isolate [28].

Gene content and *Ascaris* proteome

Earlier annotations of protein coding genes for *A. suum* draft genomes were produced by Jex [25] and Wang [27] and improved with a recent updated genome [28]—though the focus of the recent study was not on protein annotations. Here, we updated, identified, and fully annotated the 17902 protein-coding genes in the reference-based genome assembly (Table S2 and Figure S1), an annotation that can be a resource for others studying *Ascaris*. Using a custom pipeline (see methods and [31]), we classified 48% of the predicted proteome into functional groups (Figure 1A). Although the remaining 52% (9300) of the genes were classified as unknown/uncharacterized, 2515 (27%) of these appear to encode proteins that have signatures indicative of either being secreted or being membrane-bound (some with GPI anchors). To provide a more comprehensive annotation of the transcriptomes of *A. suum* and *A. lumbricoides*, we re-mapped the RNA-seq data from *A. suum* to the current gene models of *A. lumbricoides* (ALV5) (Table S2). We performed multivariate analyses of this revised RNA-seq data compilation to generate a comprehensive RNA-seq data set for differential gene expression in diverse stages/tissues (Table S2).

Phylogenetic trees derived from orthologue analyses of the predicted proteomes of ALV5 with the predicted proteomes of other nematodes across all clades indicated the similarity among the published genomes of *A. suum* [PRJNA62057 and PRJNA80881 in [25,27,28]] and *A. lumbricoides*

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[32] with ALV5 within the *Ascaris* branch (Figure 1C). The variation observed within the *Ascaris* spp. (with relatively weak bootstrap values of 0.3-0.59) is likely due to the differences in protein coding gene annotations and split genes seen in previous assemblies.

Mitochondrial genome assembly

We next took advantage of the abundant reads from the mitochondrial genome in our sequencing data (on average 7690X coverage, see Table S1) to perform *de novo* assembly of 68 complete human isolate *Ascaris* spp. mitochondrial genomes from individual worms (Table S3). These mitochondrial genomes were then annotated using sequence similarity to well-characterized and annotated mitochondrial genes.

Population structure inferred from mitochondrial CO1 gene

The mitochondrial CO1 gene has been frequently used to infer evolutionary distances between species as well as between populations [22,33–37] due to its rapid mutation rate, lack of recombination and relatively constant rate of change over time [38–40]. Previous CO1 phylogeny studies resolve *Ascaris* spp. worms into three distinct clades: clade A is predominantly comprised of worms isolated from pigs, clade B is predominantly comprised of worms isolated from humans, and clade C is from worms only isolated from pigs in Europe and Asia [22]. Interestingly, haplotype network analyses revealed that the majority of worms isolated from humans in the Kenyan villages possessed CO1 haplotypes that were consistent with infection of parasites from clade A (63/68), whereas only 6 isolates had CO1 haplotypes consistent with infection by worms from clade B (Figure S2 and Figure 2a)).

When CO1 sequences from the present study were compared against those within the *Ascaris* species complex deposited at NCBI (see Table S4 and Figure 2B)[22,23,41,42], within clade A (which appeared to contain the majority of sequences not only from Kenya but also from other localities), 7 unique haplotypes of CO1 from Kenya were identified. These appeared to be shared not only with other haplotypes from Africa, but also with those from Brazil. In contrast, clade B haplotypes appeared to be even more cosmopolitan, with the three haplotypes from Kenya not only

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

being shared with Zanzibar, but also with haplotypes from Brazil, Denmark, China and Japan. Despite the distinct clustering of haplotypes into the three typical *Ascaris* clades, there was very little genetic diversity among haplotypes within each of the clades, with the majority of haplotypes being separated by 1-4 nucleotide differences. There were greater levels of genetic divergence between clades; A and B were closer to each other while C was more distinct. Similar findings were seen with ND4, the most variable gene in the mitochondrial genome (Figure S2, Figure S3, supplemental text).

Phylogenetic analyses and population structure inferred from complete mitochondrial genomes

Forty-seven SNPs were identified in the human *Ascaris* mitochondrial genomes. Approximately a quarter of these variants were in non-coding portions of the mitochondrial genome and half were synonymous (Table S1). As with the CO1 haplotype analyses, whole mitochondrial genome analysis distinguished two clades (clade A and clade B), but there were no distinct geographically specific sub-clades seen within either clade A or B (Figure 2B and Table 2). Clade C was also produced by a single published sequence which was used for comparison. In order to assess the validity of the clades A and B representing two distinct molecular taxonomic units, and thus potentially different species, Birky's [41] 4X ratio was applied to provide a lineage specific perspective of potential species delimitation. The ratio failed to differentiate clades A and B as distinct species with $K/\Theta<4$ at 2.285 indicating *Ascaris* is one large population—further supporting the lack of differentiation into separate species (Table S5). Furthermore, there were no significant associations between mitochondrial sequence variations and other factors (e.g. village, household, time of worm collection, host) based on PERMANOVA (see methods and Table 2) after translating the phylogenetic tree into a distance matrix, suggesting not only a lack of differentiation into distinct species but also a potentially large interbreeding population of worms being transmitted between individuals and across villages.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To account for a potentially large population of interbreeding worms, analyses to detect signatures of population expansion were performed. When the global mitochondrial genome data were compared, the Tajima's D was negative and significant (Tajima's D -1.5691; P-value 0.028), indicating an excess of low frequency polymorphisms within the global data set suggesting population size expansion. Despite the Fu's F not being significant it was positive (Fu's Fs 8.5673; P-0.975) potentially indicating a deficiency in diversity as would be expected in populations that have recently undergone a bottle neck event. The same pattern was also seen in the Kenyan sequences but neither the Tajima's D nor the Fu's were significant. Although there does appear to be a signature of a recent population expansion event in both the global and Kenyan data, the lack of information on the mutation rates of *Ascaris* and other nematodes prevents the accurate estimate of such an event.

Nuclear Genome Variation in the *Ascaris* population

To quantify genetic variation in the *Ascaris* worms isolated from infected Kenyans, the nuclear genomes of the 68 individual worms were analyzed to assess intraspecific population genetic diversity, heterozygosity, and ploidy. Single nucleotide polymorphisms (SNPs) and insertion/deletions (InDels) across the nuclear genomes was assessed for the first 50 largest scaffolds, that comprised 78% of the genome (see methods). Each *Ascaris* worm was sequenced to a mean coverage depth of ~27-fold. A total of 11.15 million SNP positions were identified in the first 50 scaffolds among the *Ascaris* nuclear genomes. Approximately 25% of these variants were intergenic (Table S1). As an example, SNPs and InDels in a single *Ascaris* chromosome were plotted for two worms collected from humans in Kenya and one worm from a pig in the United States (Figure S4). The profiles and the frequency between SNPs and InDels are highly consistent within individual worms, with the ratio of InDel:SNPs frequency at ~1:7. A comparison of the variations identified between individuals infected with worms that had either *A. lumbricoides*-like or *A. suum*-like mitochondrial genomes illustrates that most of the differences appear to be random variations, and there do not appear to be major differences between *A. lumbricoides*-like and *A.*

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

suum-like worms. A total of 1.79 million SNPs were private, or unique to individual strains, presumably representing genetic drift. Of the remaining 9.3 million SNPs, ~32% of these variant positions were present in less than 5 isolates indicating that the *Ascaris* genomes sequenced are ~1% polymorphic among the major alleles circulating within the species complex.

Population structure inferred from nuclear genomes

To investigate the evolutionary pressures that account for the high SNP diversity found among the 68 sympatric isolates, the ploidy, degree of heterozygosity (*He*) and allelic diversity was determined. Worms were disomic, with little to no evidence of aneuploidy (Figure S5). The vast majority (>98%) of SNP positions were biallelic, and each isolate had, on average, 2.3 million variant positions, of which approximately 60% were heterozygous SNPs (Table S6). SNP density was determined in 10kb windows for each worm against the reference ALV5 isolate and a patchy, mosaic pattern was resolved. SNP density was structured within the genome, with scaffolds being either SNP poor or SNP dense. For example, Algv5r020 was SNP dense whereas Algv5r019x was SNP poor. In other scaffolds, alternating SNP poor and SNP dense regions were defined within the contig, with distinct transition points, see for example the first half of Algv5b02, the last quarter of Algv5b05, or the middle of Algv5r021x (Figure 3A). In those regions where SNP density was low, the Tajima D statistic was net negative, indicating that allele frequencies within these regions were structured and more limited.

Genome-wide, homozygous SNP regions were found to be unevenly distributed, with some scaffolds possessing long runs of homozygosity, see for example Algv5b02, Algv5r009x, Algv5r013x, Algv5r014x, Algv5r018x, Algv5r019x, Algv5r027x (depicted by solid blue in Figure 3B), and these regions were net negative by the Tajima D test. Conversely, heterozygous SNPs were less structured and appeared randomly distributed throughout the genome (Figure 3B). Overall, three genetic types were resolved by this analysis: in each genome there existed SNP-poor homozygous regions (colored blue) or SNP dense regions, that either possessed homozygous alternate SNPs (also colored blue) or heterozygous SNPs (colored in “red” or “yellow” blocks

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

depending on the density of heterozygous SNPs resolved in each 10kb block: one haplotype was similar to ALV5 and the other was different). Only one isolate (119_3) was heterozygous genome-wide, and this track is depicted as "red" across all scaffolds in the Circos plot (Figure 3B).

Population genetic structure of Kenyan *Ascaris* worm isolates

A phylogenetic tree constructed using genome wide SNPs with at least 10x coverage (11.15 million phased SNPs total) from 69 Ascaris strains, including the *A. suum* reference genome, established that the Kenyan isolates were more similar to each other than they were to the *A. suum* reference genome, which had many more private SNPs (Figure 4A). Notably, the nuclear genomes from the isolates that possessed *A. lumbricoides*-like mitochondrial genomes did not clade separately, indicating that the nuclear genomes were incongruent with the mitochondrial genomes, and likely recombinant. A co-ancestry heatmap was generated among the sympatric *Ascaris* isolates, and this analysis divided the genome into discrete segments and clustered samples along the diagonal based on the greatest number of shared ancestral blocks using the nearest neighbor algorithm from fineSTRUCTURE. The *Ascaris* genomes resolved as thirteen clusters that possessed high frequency nearest-neighbor, or shared ancestry, relationships. In contrast, the *A. suum* reference genome and strain 119_3 were anomalous, likely the result of their excess heterozygosity due in part to elevated numbers of private SNPs. Notably, 9 isolates did not coalesce into a cluster with shared ancestry. Closer examination of these strains indicated that their phased genomes possessed limited allelic diversity and were highly recombinant (Figure 4B). This genetic mosaicism was readily resolved by fluctuating intra-scaffold genealogies established using a sliding-window neighbor-joining topology that identified regions with incongruent tree topologies. See for example the trees generated at the scaffolds ALgV5b01, ALgV5b02, and ALgV5r001. Indeed, the pairwise SNP and F_{ST} estimates for these strains identified segments where SNP density was low, but F_{ST} was elevated with respect to neighboring segments (see block in ALgV5b02) and the most parsimonious explanation for these results is that recombination of a limited number of distinct alleles had occurred in the regions of increased F_{ST} (Figure 4B and 4C).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To estimate the number of supported ancestries (K) that could be resolved in the *Ascaris* genomes sequenced, we calculated the Dunn index, which supported 3-6 ancestral populations (Figure 4D). A gradual increase in the Dunn Index after K = 6 was observed for an ancestral population size between 2 and 15 (Figure 4D and Figure S6). We next used POPSICLE to calculate the number of clades present within each 10kb sliding window. Local clades were represented with a different color and painted across the genome to resolve ancestry. The SNP diversity plots across the 68 isolates identified 3 major "parentage blocks" that were resolved as belonging to ALV5 or were genetically distinct with either both haplotypes sharing the alternate parent (homozygous alternate), or were heterozygous between the two parental haplotypes for the majority of the isolates (Figure 4E, middle Circos plot. Color hues cyan, orange, aqua).

To visualize such shared ancestry across the different *Ascaris* strains at chromosome resolution, a color hue representing a local genetic "type" present was assigned and integrated to construct haplotype blocks across each chromosome for the ancestries present. Chromosome painting based on shared ancestry revealed a striking mosaic of large haplotype blocks of different admixed color hues, consistent with limited genetic recombination between a low number of parentage haplotypes. These admixture patterns were readily visualized by shared color blocks between different strains across entire scaffolds including AlgV5R019X (Figure 5A) and AlgV5R027X (Figure 5B). In low complexity regions such as the left portion of contig ALgV5R019X, only three major haplotypes were resolved (Figure 5A). Strikingly, within each of the 6 clades resolved, all strains showed a limited, mosaic fingerprint of introgressed sequence blocks indicating that recombination has shaped the population genetic structure among the *Ascaris* isolates sequenced. Evidence for both segregation and recombination were evident. For example, isolates 1107E_1 and 2110F_2 shared the same chromosome at ALgV5R019X, but entirely different chromosomes at ALgV5R027X, whereas isolates 107_1, 108_1 and 2110F_2 were identical except at the subtelomeric end of ALgV5R19X. In this region two admixture blocks were resolved; 107_1 and 2110F_2 remained similar to each other but 108_1 now possessed a sequence

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

block that was shared with isolate 119_3. This extensive chimeric pattern in chromosome painting also closely resembled the genome-wide hierarchy tree (Figure 5A). The data support a model in which the isolates are genetic recombinants between *A. suum* and *A. lumbricoides* that are predominantly inbreeding.

Geographic and demographic correlates of genetic similarity

To examine clustering of worms in similar human hosts, we statistically compared genetic variation within groups (such as within a village) versus between groups (such as between villages). We found significant genetic separation between worms in different villages (using Adonis vegan in R), but not between worms from different countries (Table 2). This suggests genetic diversity is present in the population of *Ascaris* in these Kenyan villages, which is similar to the diversity of populations of *Ascaris* around the world. It also suggests that a high proportion of *Ascaris* transmission may occur within villages in this Kenyan setting. There was no evidence from this analysis that the 13 worms collected three months after albendazole treatment were any different than the worms collected prior to albendazole treatment (Table 2).

To expand on our observations, that genetically similar worms are found around the world, but that similar worms cluster within a village based on our nuclear SNPs data, we plotted genetic distances against geographic distances. Surprisingly, we found no significant correlations between genetic and geographic distance, neither across all five studied villages nor within the two most heavily parasitized villages (Figure S7 and supplemental text).

Discussion

In this study, we generated a high-quality reference genome from a single worm presumed to be human *A. lumbricoides*. Our comparative phylogenomic analyses of this new *Ascaris spp.* genome against existing draft genomes of *A. lumbricoides* and *A. suum* suggest that *A. suum* and *A. lumbricoides* form a genetic complex that is capable of interbreeding that has apparently undergone a recent worldwide, multi-species *Ascaris* population expansion.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Our phylogenetic analysis on the complete mitochondrial genomes (from 68 worms collected from human hosts in Kenya and other available sequences) suggests that the worms collected in Kenya mirrors the separation into clade A (worms from pigs in non-endemic regions and humans in endemic regions) and clade B (worms from humans and pigs from endemic and non-endemic regions) described elsewhere [22]. It is likely that worms in both of these clades are being transmitted from human to human, as pig husbandry is rare in this area of Kenya. Patterns may differ by locality, and it is possible that some of the pig-associated (*A. suum*-like) worms circulating in this human population in Kenya were acquired, perhaps generations ago, by humans who lived in closer proximity to pigs. It is also possible that these worms were acquired from non-human primates [42], or some other *Ascaris* host, rather than from pigs.

However, the SNPs across the whole nuclear *Ascaris* genome provide significantly greater power in understanding *Ascaris* speciation Importantly, our nuclear genome SNP analysis suggest that the 68 Kenyan *Ascaris* are distributed across multiple clades in a phylogeny based on the nuclear genomes. Overall, data from our study and other studies are consistent with a pattern where hybrid genotypes in *Ascaris* populations were observed [11,22,43] Our study represents one of the most detailed accounts of mito-nuclear discordance in nematodes echoing patterns seen in another human nematode *Onchocerca volvulus* [44]. The data in our current study shows the occurrence of distinct mitochondrial lineages that could be evidence of early stages of species differentiation. The admixture seen within the nuclear genome, however, appears to disrupt the establishment of defined molecular speciation barriers between the different *Ascaris* lineages. Such patterns have been recorded in other parasites not only in *O. volvulus* [44], but also in the parasitic blood fluke *Schistosoma* [45] and the protist *Leishmania* [46]. Each of the studies have implicated definitive hosts in the movement of parasites between otherwise isolated populations allowing interbreeding to take place. It is most likely the historical movement of humans and their domesticated livestock that has mediated the transport of *Ascaris* between localities, allowing for

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

372 extensive interbreeding within the nuclear genomes and the discordance observed between the
373 mitochondrial and nuclear genomes in our study.

374 At a more local scale, the insights into the human transmission dynamics of *Ascaris* showing
375 clustering both within an individual and in villages suggest that villages are appropriate units for
376 interventions and that people are infected with multiple eggs from a single source. These findings
377 are in line with clustering at the village level found in Guatemala [47] and at the sub-village level in
378 Nepal [48], but not in line with the lack of small-scale geographical structuring found in Denmark,
379 Zanzibar and Uganda [49–51]. Differences could be a result of different patterns in human and
380 livestock movement [20].

381 Although the current genome is, by far, the most continuous assembly for *Ascaris*, it is not a
382 full chromosome assembly due largely to repetitive sequences, in particular 120 bp tandem repeat
383 clusters and long stretches of subtelomeric repeats. Thus, it is possible that mis-assembly in some
384 scaffolds has increased the frequency of mosaicism detected. It is for this reason that the
385 comparative analyses on the nuclear genome was restricted to the largest 50 scaffolds, most of
386 which are at chromosomal resolution, with only minor localized variation due to the repeat clusters.
387 In these high confident scaffolds, large haplotype blocks possessing either *A. suum, A. lumbricoides*
388 or both parental haplotypes (heterozygous) were readily resolved indicating that the genetic
389 mosaicism observed could not be solely attributed to genome mis-assembly. Ultimately, future
390 studies using ultralong PacBio and/or Nanopore sequencing combined with chromosome
391 conformation capture (Hi-C) techniques will improve the genome to full chromosome assembly to
392 more accurately resolve the true extent to which recombination has impacted the population genetic
393 structure of the *Ascaris* species genetic complex.

394 The finding that *A. suum* and *A. lumbricoides* form a genetic complex has important public
395 health implications. Reduced treatment efficacy is not currently a common issue in *Ascaris*
396 infections among humans or pigs [52–54], though low efficacy of benzimidazoles is an issue for
397 *Trichuris trichiura* in humans [55–57] and various intestinal nematodes of veterinary importance [58–60].

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Extensive albendazole use in either human or pig populations could lead to resistance in both populations, if cross-species infections are common and produce fertile offspring. This study suggests that research and public health interventions targeting *A. lumbricoides* and *A. suum* should be more closely integrated, and that extensive work done by the veterinary research community may be highly relevant to mass deworming campaigns that seek to improve human health.

The similarity between *Ascaris* from different countries and from different vertebrate hosts suggests that *Ascaris* infection has spread rapidly around the world, leaving little time for it to differentiate. Taken together, these finding have very important implications for parasite control and elimination efforts that only focus on mass deworming of humans for *Ascaris*. The ability of pig-associated worms to become endemic in human populations indicates that a one-health approach may be necessary for the control of *Ascaris*.

Online Methods

Worm collection

Worms were expelled as part of a larger study in rural western Kenya described previously [61,62]. Worms collected from study participants in five villages (map shown in [61]) following treatment with 400 mg albendazole were isolated, washed, labelled and stored frozen (-15 C). They were transported from Bungoma to Kisumu, where they were subsequently stored at the KEMRI-CDC offices until they were shipped to the NIH (Bethesda, MD, USA) on dry ice.

DNA extraction and sequencing

A modified DNA extraction method was developed (see Supplemental Methods and Table S3). For the five germline samples, DNA was extracted from the uterus, oviduct or ovary of the worms. For the remaining samples, DNA was extracted from somatic tissue: the body wall or the intestine.

Paired-End Genome Libraries – Sixty-eight *A. lumbricoides* DNA samples were sequenced using Illumina HiSeq 2500 (www.illumina.com) short-read paired-end sequencing. DNA was quantified by UV Spec and Picogreen. A 100 ng of DNA based on picogreen quantification was used as

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

template for NGS library preparation using the TruSeq Nano DNA Sample library prep kit without modification. Primer-dimers in the libraries were removed by additional AMPure beads purification. Sequencing was performed to obtain a minimum genomic depth of 20X coverage for each sample.

Mate-Pair Genome Libraries – Two samples were selected for mate-pair sequencing, based on the quality of the DNA preparation. Three independent DNA isolations (corresponding to what region of the worm or what is the sample for DNA isolation) from specimen "119_2.3" were combined to obtain one µg DNA input. The mate-pair libraries were generated using the Nextera Mate Pair Library Prep Kit, following the gel-free method with the only modification that M-270 Streptavidin binding beads were used instead of M-280 beads. The libraries were amplified for 15 cycles given the low DNA input going into the circularization phase. The mate-pair fragment size averaged 6 kb with a range of 2-10 kb fragments.

Assembly and annotation of A. lumbricoides reference genome

The *A. lumbricoides* germline genome assembly was constructed using the *A. suum* genome as a reference. Briefly, sequencing reads from a single *A. lumbricoides* worm (libraries #8457, #8458 and #8778) were mapped to the *A. suum* germline genome assembly [28] using BWA[63] to generate BAM and MPILEUP alignment files. The MPILEUP files were processed with a PERL script that replaced all variation sites in the reference genome with the highest allele frequencies in the *A. lumbricoides* sample. *A. suum* genomic regions that represent < 5X of *A. lumbricoides* reads coverage were excluded from the assembly. We further polished the genome with additional Illumina sequencing reads using Pilon and its default parameters [64]. The *A. lumbricoides* genome was annotated using the gene models built for *A. suum*, using the annotation transfer tool RATT [65]. The protein coding regions were defined using TransDecoder (https://github.com/TransDecoder/TransDecoder/wiki). To evaluate the gene expression across all stages, we utilized previous RNAseq data from the developmental stages [27,28], re-mapped the SRA from adult males, females, L3 and L4 stages [25] to the current gene models, and quantified the

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expression using tophat and cufflinks. The re-mapped reads, analyzed by JMP Genomics (SAS) across all the stages and based on the principal component analyses (Figure 1B), were grouped as adult male, adult female, L1, L2, L3 (egg L3, liver L3 and lung L3), L4, carcass, muscle, intestine, embryonic (zygote1, zygote2, zygote3, zygote4, 24h, 46h, 64h, 96h, 5d, 7d), ovaries (female mitotic region, female early pachytene, female late pachytene, female diplotene and oocyte) and testis (male mitotic region, spermatogenesis, post meiotic region, seminal vesicles and spermatids). Proteome and comparative genomics analyses were done using an in-house pipeline [66]. Automated annotation of proteins was done as described earlier [31] and based on a vocabulary of nearly 290 words found in matches to various databases, including Swissprot, Gene Ontology, KOG, Pfam, and SMART, Refseq-invertebrates and a subset of the GenBank sequences containing nematode protein sequences, as well as the presence or absence of signal peptides and transmembrane domains. Signal peptide, SecretomeP, transmembrane domains, furin cleavage sites, and mucin-type glycosylation were determined with software from the Center for Biological Sequence Analysis (Technical University of Denmark, Lyngby, Denmark) [67–69]. Classification of kinases was done by Kinannote [70]. Interproscan [71] analyses were done using the standalone version 5.34. Allergenicity of proteins were predicted by Allerdictor [72], FuzzyApp [73] and AllerTOP [74]. Genes that had blast scores <30% of max possible score (self-blast) in other non-Ascaris nematodes with an e-value greater than 1E-05 were considered as 'unique'. The orthologues of predicted proteome of ALV5 across the publicly available nematode genomes (*Ancylostoma caninum* [32], *Ancylostoma ceylanicum* [32,75], *Ancylostoma duodenale* [32], *Ascaris lumbricoides*[32], *Ascaris suum* [25,27,28], *Brugia malayi* [76], *Caenorhabditis elegans* [77], *Dirofilaria immitis* [78], *Loa loa* [79,80], *Necator americanus* [81], *Onchocerca volvulus* [31], *Strongyloides ratti* [82], *Strongyloides stercoralis* [83], *Toxocara canis* [32,84], *Trichinella spiralis* [85,86], *Trichuris trichiura* [87], *Wuchereria bancrofti* [32,88]) were analyzed using OrthoFinder [89]. The estimated phylogenetic tree generated was graphed using FigTree v1.4.

Further manual annotation was done as required. The data were mapped into a hyperlinked Excel spreadsheet as previously described [90], available in Table S2.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Read mapping and SNP analysis for whole genome sequences

The Illumina paired end sequence reads of the 68 *Ascaris* whole genomes were trimmed by removing any adapter sequences with CutAdapt v1.12 [91], then low quality sequences were filtered and trimmed using the FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Remaining reads were then ref-mapped to the *A. lumbricoides* genome ALV5 reference genome (described in this paper) using either Bowtie2 v2.2.9 [92], with very sensitive, no-discordant, and no-mixed settings or using the Burrows-Wheeler Aligner (BWA, v0.7.9) [63] mem in default parameters and then converted into a bam file for sorted with SAMtools [93]. Sorted reads were soft-clipped and marked-duplicated using Picard-1.8.4 (http://broadinstitute.github.io/picard). Single nucleotide polymorphisms (SNPs) were obtained using SAMtools [93] and BCFtools [94] using the mpileup function and –ploidyfile features and taking chromosomal ploidies into account. SNPs were also determined using Genome Analysis Toolkit (GATK) [95]. SNPs were called by GATK Haplotype Caller with a read coverage ≥ 10x, a Phredscaled SNP quality of ≥30. Mapping statistics were generated in Perl and Awk.

Ploidy determination

The ploidy of each isolate was calculated using AGELESS software (http://ageless.sourceforge.net/) by dividing the chromosomes into 10kb sliding windows and averaging the coverage within each window. The windows with zero coverage were not included in any further analyses due to sequencing noise or repeat regions [96].

Genetic diversity

SNPs, pi [97], TajimaD [98], and F_{ST} [99] values were calculated using VCFtools [100] in 10 kb sliding windows and plotted using either Circos [101] or ggbio (http://bioconductor.org/packages/release/bioc/html/ggbio.html) and VariantAnnotation (http://bioconductor.org/packages/release/bioc/html/VariantAnnotation.html) R packages (v. 3.1.0, URL http://www.R-project.org). The proportions of heterozygous and homozygous SNPs were

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

501 estimated in 10kb sliding windows using custom Java scripts to generate histogram plots in Circos
502 [101]. Red and blue colors indicate the presence of 90% or more heterozygous and homozygous SNPs
503 respectively whereas yellow color was assigned otherwise.

504 *Co-ancestry heatmap*

505 The SNP data (VCF file) was first phased accurately to estimate the haplotypes using SHAPEIT [102]
506 after keeping only biallelic SNPs and loci with less than 80% missing data. Co-ancestry heatmaps
507 were generated using the linkage model of ChromoPainter [103] and fineSTRUCTURE
508 (http://www.paintmychromosomes.com) based on the genome-wide phased haplotype data. For
509 fineSTRUCTURE (version 0.02) [103], both the burn-in and Markov Chain Monte Carlo (MCMC)
510 after the burn-in were run for 1000 iterations with default settings. Inference was performed twice
511 at the same parameter values.

512 *Population genetic structure*

513 Population genetic structure was constructed using POPSICLE [104] by comparing strains against the
514 reference sequence ALV5 in 10 kb sliding windows with the number of cluster K=1 to 15 and then
515 use the Dunn index [99] to calculate the optimal number of clusters. After calculating the optimal
516 number of clusters, POPSICLE assigned each block to the existing or new clades depending on
517 population structure of each strain and the ancestral state of each block followed by painting in
518 Circos plot [101] with color assignment based on number of clusters.

519 *Construction of phylogenetic trees*

520 In order to determine the phylogenetic relationship between samples, we selected 19005 base
521 positions where variants were detected in a representative sample vs the reference (ALV5), and
522 where each sample had at least 20x coverage for each locus. Using this list, the base calls for each
523 sample were pooled together to generate a single multi sequence fasta file.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Next, both maximum likelihood (ML) trees and bootstrap (BS) trees were generated with a final "best" tree generated from the best scoring ML and BS trees using RAxML v8.2.10 [105]. The tree was visualized in FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

Permutational Multivariate Analysis of nuclear phylogeny

Similarity within and between worms from different villages, households, people and time-points was analyzed based on the distance matrix of the patristic distances from the phylogenetic tree described above, using permutational multivariate analysis of variance (Adonis Vegan in R). The distance matrix underlying the phylogenetic tree was analyzed in order to measure the significance and contribution of different factors to variance between samples. Each factor (village, household, host, time-point, human age group and worm age/size group) was analyzed both separately and sequentially. The sequence chosen was ordered based on significance of each factor when tested individually. Since multiple groupings were considered using the same dataset, multiple comparison corrections were applied. Sample sizes and descriptions of each group are shown in Table 2. Worms with recorded lengths were split into three groups of equal number. This was in order to test whether larger worms (using worm length as a very approximate proxy for age, where older worms might have survived at least one round of ALB treatment) were any different from the newly acquired younger worms. Similar methods were used to analyze the mitochondrial phylogeny along the same groupings.

Mitochondrial genome assembly

We assembled mitochondrial genomes using a *de novo* approach from 68 individual *Ascaris* genomes. For each individual, the *Ascaris* mitochondrial reads in the total DNA sequencing were identified by mapping the *Ascaris* reads to the *A. suum* reference mitochondrial genome (GenBank accession: NC_001327). Adaptor sequences were trimmed prior to *de novo* assembly. To reduce the complexity of the *de novo* assembly, we randomly sampled 1,000x reads from each individual (the use of higher read coverage often resulted in fragmented scaffolds) and assembled these reads using the SPAdes assembler [106] with continuous k-mer extension from K=21 to the maximum k-

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mer allowed (average extended k-mer size = 91). The assembled scaffolds were corrected with the built-in tool in SPAdes to reduce potential assembly artifacts. Next, the assembled scaffolds were aligned to the *A. suum* mitochondrial reference genome using BLAST, the order of the scaffolds was adjusted, and they were joined into a single scaffold. Finally, the gaps in the scaffold were filled using GapFiller [107] using mitochondrial reads from the same individual to generate a complete mitochondrial genome. Using the same method, we also *de novo* assembled another five *A. suum or A. lumbricoides* mitochondrion genomes from previous studies (see Table S7).

Analysis of mitochondrial genomes

In order to assess overall evolutionary relationships across the complete mitochondrial genomes, we aligned the genomes using Clustal W and phylogenetic trees constructed using RaxML under the conditions of the general time reversible model (GTR) as described above for the whole genome SNP alignment. Subsequent tree files were formatted in FigTree and MEGA v7. The variation in nucleotide diversity across the mitochondrial genome was measured using sliding window analyses, with a window of 300 bp and a step of 50bp, using DNAsp v6 [108]. In order to assess the validity of potential species groupings in the ML phylogenetic tree the Birky [41] X4 ratio was applied to the alignment of the complete mitochondrial genomes including both samples from Kenya and published mitochondrial reference genomes from Tanzania, Uganda, China, USA, Denmark and the UK. The X4 ratio method of species delimitation compares the ratio of mean pairwise differences between two distinct clades (K) and the mean pairwise differences within each of the clades being compared (Θ). It is considered that if K/Θ>4 this is indicative of the two clades representing two distinct species. Owing to the fact that two clades are being compared there will be two separate values of Θ, as per recommendations of Birky [41], the larger Θ value is used to perform the final ratio calculation as this will provide a more conservative result which ultimately will be less likely to provide a false positive result.

Due to the extensive use of mitochondrial genome data in population genetic analyses of *Ascaris*, several analyses were performed to identify the effect of any population level processes that may be

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

affecting the diversity of the parasites within Kenya. Initially, diversity indices were calculated for each of the genes within the mitochondrial genome across the entire Kenyan data set as well as considering the mitochondrial genome as a whole. In order to account for the diversity within the genic regions, we removed non-coding and tRNA sequences for these analyses. To provide a genealogical perspective of population structure of the Kenya *Ascaris* isolates, we constructed the most parsimonious haplotype network based on the protein coding sequences using the TCS algorithm as implemented in PopArt [109]. Further population genetic analyses were also performed to detect the occurrence of selection on the protein coding genes of the mitochondrial genome and if there were any major departures from neutrality. Standard dN/dS ratios were performed to identify the presence of positive selection where both measures equate to 1 = neutral, >1 = positive selection, <1 = purifying selection. Both Tajima's D and Fu's *Fs* were calculated to identify any substantial departure from neutrality which could be indicative of population expansion events (Table S8). All described analyses were performed using DNAsp6 [108]. As both CO1 and ND4 have been used in the past for epidemiological studies, single gene phylogenies were also constructed as described previously for comparison against the whole mitochondrial genome phylogeny (Figure S3 and supplementary methods).

Owing to the extensive use of the CO1 for epidemiological studies the gene was extracted from the complete mitochondrial genomes of Kenya and compared to all other available *Ascaris lumbricoides* and *Ascaris suum* CO1 sequences housed by NCBI representing populations from across the globe. Haplotype network analyses was performed to produce the parsimonious network using TCS as implemented through PopArt [109]. This provided a genealogical perspective of population structure and allowed genetic connectivity between the Kenyan samples and other geographical isolates to be assessed.

Acknowledgements

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We would like to thank the school children, schoolteachers, and Bungoma administrators for their support. We would like to extend special thanks to all the members of the study team: Bungoma County Hospital, Siangwe, Siaka, Sang'alo, Nasimbo and Ranje village administrators and Community Health Workers. Particular thanks to Dr. Charles S. Mwandawiro, Prof. Sammy Njenga, and Dr. Jimmy H. Kihara (KEMRI), and Dr Simon J. Brooker (BMGF) for making the fieldwork possible in Kenya, and for their invaluable scientific and logistical advice.

Data availability

Data are available under the National Center for Biological Information (NCBI) BioProject numbers; PRJNA511012 and SRA submission SUB4949491 for sequencing data, and PRNJA515325 for the genomic assembly.

Links to all genome assemblies are available at: All, de novo, semi-de novo (V1), V2, V3, V4, V5 and mitochondrial.

Funding

This work was supported in part by the Division of Intramural Research (DIR) of the National Institute of Allergy and Infectious Diseases, NIH and NIH grants to R.E.D. (AI114054) and J.W. (AI125869). Fieldwork was supported by a grant from the Bill and Melinda Gates Foundation to the London Centre for Neglected Tropical Disease Research and the KEMRI Wellcome Trust. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

AVE, RMA, RED, JW, TBN conceived the study; SG, ED, JW, ED, SFP, RED performed sequencing and genome analyses; SPL, JPW, AK, MEG performed population biology and evolutionary biology expertise and analyses; AVE, SK, RGO, RMA oversaw the collection of

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

material in Kenya, SB, SG, ED, JW provided protein/gene content analyses, AVE, SPL, SB, AK, ED MED, RED, JW, TBN wrote the manuscript.

Ethics approval

This study was approved by the Ethics Review Committee of the Kenya Medical Research Institute (Scientific Steering Committee protocol number 2688) and the Imperial College Research Ethics Committee (ICREC_ 13_1_15). Informed written consent was obtained from all adults and parents or guardians of each child. Minor assent was obtained from all children aged 12–17. Anyone found to be infected with any STH was treated with 400 mg ALB during each phase of the study, and all previously-untreated village residents were offered ALB at the end of each study phase.

Consent for publication

Individuals consented to the publication of their results, without any patient identifying information.

Competing interests

The authors declare that they have no competing interests. RMA was a Non-Executive Director of GlaxoSmithKline (GSK) during the period of worm collection in Kenya. GSK played no role in the funding of this research or this publication.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Pullan, R. L., Smith, J. L., Jasrasaria, R. & Brooker, S. J. Global numbers of infection and disease burden of soil transmitted helminth infections in 2010. *Parasit. Vectors* **7,** 37 (2014).
2. Hotez, P. J. *et al.* Helminth infections: the great neglected tropical diseases. *J. Clin. Invest.* **118,** 1311–1321 (2008).
3. Montresor, A. *et al.* Preventive chemotherapy and the fight against neglected tropical diseases. *Expert Rev Anti Infect Ther* **10,** 237–242 (2012).
4. Brooker, S. Estimating the global distribution and disease burden of intestinal nematode infections: adding up the numbers--a review. *Int. J. Parasitol.* **40,** 1137–1144 (2010).
5. Bundy, D. A. P. *et al.* in *Child and adolescent health and development* (eds. Bundy, D. A. P., Silva, N. de, Horton, S., Jamison, D. T. & Patton, G. C.) (The International Bank for Reconstruction and Development / The World Bank, 2017). doi:10.1596/978-1-4648-0423-6/pt3.ch13
6. Becker, S. L. *et al.* Toward the 2020 goal of soil-transmitted helminthiasis control and elimination. *PLoS Negl. Trop. Dis.* **12,** e0006606 (2018).
7. Nejsum, P., Betson, M., Bendall, R. P., Thamsborg, S. M. & Stothard, J. R. Assessing the zoonotic potential of Ascaris suum and Trichuris suis: looking to the future from an analysis of the past. *J Helminthol* **86,** 148–155 (2012).
8. Anderson, T. J. & Jaenike, J. Host specificity, evolutionary relationships and macrogeographic differentiation among Ascaris populations from humans and pigs. *Parasitology* **115 (Pt 3),** 325–342 (1997).
9. Nejsum, P. *et al.* Ascariasis is a zoonosis in denmark. *J. Clin. Microbiol.* **43,** 1142–1148 (2005).
10. Dutto, M. & Petrosillo, N. Hybrid ascaris suum/lumbricoides (ascarididae) infestation in a pig farmer: a rare case of zoonotic ascariasis. *Cent. Eur. J. Public Health* **21,** 224–226 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

674 11. Betson, M., Nejsum, P., Bendall, R. P., Deb, R. M. & Stothard, J. R. Molecular epidemiology
675 of ascariasis: a global perspective on the transmission dynamics of Ascaris in people and
676 pigs. *J. Infect. Dis.* **210,** 932–941 (2014).
677 12. Miller, L. A. *et al.* Ascariasis in humans and pigs on small-scale farms, Maine, USA, 2010-
678 2013. *Emerging Infect. Dis.* **21,** 332–334 (2015).
679 13. Anderson, T. J. Ascaris infections in humans from North America: molecular evidence for
680 cross-infection. *Parasitology* **110 (Pt 2),** 215–219 (1995).
681 14. Zhu, X., Chilton, N. B., Jacobs, D. E., Boes, J. & Gasser, R. B. Characterisation of Ascaris
682 from human and pig hosts by nuclear ribosomal DNA sequences. *Int. J. Parasitol.* **29,** 469–
683 478 (1999).
684 15. Peng, W. & Criscione, C. D. Ascariasis in people and pigs: new inferences from DNA
685 analysis of worm populations. *Infect. Genet. Evol.* **12,** 227–235 (2012).
686 16. Takata, I. Experimental infection of man with Ascaris of man and the pig. *Kitasato Arch Exp*
687 *Med* **23,** 151–9; English transl, 49 (1951).
688 17. Betson, M. & Stothard, J. R. Ascaris lumbricoides or Ascaris suum: What's in a Name? *J.*
689 *Infect. Dis.* **213,** 1355–1356 (2016).
690 18. da Silva Alves, E. B. *et al.* Ascaris lumbricoides, Ascaris suum, or "Ascaris
691 lumbrisuum"?Multiple Exposures to Ascaris suum Induce Tissue Injury and Mixed Th2/Th17
692 Immune Response in Mice. *J Infect Dis* **213,** 1355 (2016).
693 19. Leles, D., Gardner, S. L., Reinhard, K., Iñiguez, A. & Araujo, A. Are Ascaris lumbricoides
694 and Ascaris suum a single species? *Parasit. Vectors* **5,** 42 (2012).
695 20. Betson, M., Nejsum, P. & Stothard, J. R. in *Ascaris: the neglected parasite* 265–285
696 (Elsevier, 2013). doi:10.1016/B978-0-12-396978-1.00010-0
697 21. Peng, W., Anderson, T. J., Zhou, X. & Kennedy, M. W. Genetic variation in sympatric
698 Ascaris populations from humans and pigs in China. *Parasitology* **117 (Pt 4),** 355–361
699 (1998).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

22. Cavallero, S., Snabel, V., Pacella, F., Perrone, V. & D'Amelio, S. Phylogeographical studies of Ascaris spp. based on ribosomal and mitochondrial DNA sequences. *PLoS Negl. Trop. Dis.* **7,** e2170 (2013).

23. Zhou, C., Li, M., Yuan, K., Hu, N. & Peng, W. Phylogeography of Ascaris lumbricoides and A. suum from China. *Parasitol. Res.* **109,** 329–338 (2011).

24. Nejsum, P. *et al.* Ascaris phylogeny based on multiple whole mtDNA genomesHigh prevalence of Strongyloides stercoralis in school-aged children in a rural highland of north-western Ethiopia: the role of intensive diagnostic work-up. *Infect Genet Evol* **48,** 4–9 (2017).

25. Jex, A. R. *et al.* Ascaris suum draft genome. *Nature* **479,** 529–533 (2011).

26. Wang, J. *et al.* Deep small RNA sequencing from the nematode Ascaris reveals conservation, functional diversification, and novel developmental profiles. *Genome Res.* **21,** 1462–1477 (2011).

27. Wang, J. *et al.* Silencing of germline-expressed genes by DNA elimination in somatic cells. *Dev. Cell* **23,** 1072–1080 (2012).

28. Wang, J. *et al.* Comparative genome analysis of programmed DNA elimination in nematodes. *Genome Res.* **27,** 2001–2014 (2017).

29. Wang, J. & Davis, R. E. Programmed DNA elimination in multicellular organisms. *Curr. Opin. Genet. Dev.* **27,** 26–34 (2014).

30. Streit, A., Wang, J., Kang, Y. & Davis, R. E. Gene silencing and sex determination by programmed DNA elimination in parasitic nematodes. *Curr. Opin. Microbiol.* **32,** 120–127 (2016).

31. Cotton, J. A. *et al.* The genome of Onchocerca volvulus, agent of river blindness. *Nat. Microbiol.* **2,** 16216 (2016).

32. International Helminth Genomes Consortium. Comparative genomics of the major parasitic worms. *Nat. Genet.* **51,** 163–174 (2019).

33. Springer, M. S. *et al.* Mitochondrial versus nuclear gene sequences in deep-level mammalian

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

phylogeny reconstruction. *Mol. Biol. Evol.* **18,** 132–143 (2001).

34. Zardoya, R. & Meyer, A. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. *Mol. Biol. Evol.* **13,** 933–942 (1996).

35. Zou, Y. *et al.* Determining geographical variations in Ascaris suum isolated from different regions in northwest China through sequences of three mitochondrial genes. *Mitochondrial DNA A DNA Mapp. Seq. Anal.* **28,** 411–415 (2017).

36. Nejsum, P. *et al.* Ascaris phylogeny based on multiple whole mtDNA genomes. *Infect. Genet. Evol.* **48,** 4–9 (2017).

37. Wiens, J. J., Kuczynski, C. A. & Stephens, P. R. Discordant mitochondrial and nuclear gene phylogenies in emydid turtles: implications for speciation and conservation. *Biological Journal of the Linnean Society* **99,** 445–461 (2010).

38. Brown, W. M., George, M. & Wilson, A. C. Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **76,** 1967–1971 (1979).

39. Giles, R. E., Blanc, H., Cann, H. M. & Wallace, D. C. Maternal inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **77,** 6715–6719 (1980).

40. Harrison, R. G. Animal mitochondrial DNA as a genetic marker in population and evolutionary biology. *Trends Ecol. Evol. (Amst.)* **4,** 6–11 (1989).

41. Birky, C. W. Species detection and identification in sexual organisms using population genetic theory and DNA sequences. *PLoS One* **8,** e52544 (2013).

42. Nejsum, P., Bertelsen, M. F., Betson, M., Stothard, J. R. & Murrell, K. D. Molecular evidence for sustained transmission of zoonotic Ascaris suum among zoo chimpanzees (Pan troglodytes). *Vet. Parasitol.* **171,** 273–276 (2010).

43. Jesudoss Chelladurai, J. *et al.* Molecular epidemiology of ascaris infection among pigs in iowa. *J. Infect. Dis.* **215,** 131–138 (2017).

44. Choi, Y.-J. *et al.* Genomic diversity in Onchocerca volvulus and its Wolbachia endosymbiont. *Nat. Microbiol.* **2,** 16207 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

752 45. Lawton, S. P., Bowen, L. I., Emery, A. M. & Majoros, G. Signatures of mito-nuclear
753 discordance in Schistosoma turkestanicum indicate a complex evolutionary history of
754 emergence in Europe. *Parasitology* **144,** 1752–1762 (2017).
755 46. Kato, H. *et al.* PCR-RFLP analyses of Leishmania species causing cutaneous and
756 mucocutaneous leishmaniasis revealed distribution of genetically complex strains with hybrid
757 and mito-nuclear discordance in Ecuador. *PLoS Negl. Trop. Dis.* **13,** e0007403 (2019).
758 47. Anderson, T. J., Romero-Abal, M. E. & Jaenike, J. Mitochondrial DNA and Ascaris
759 microepidemiology: the composition of parasite populations from individual hosts, families
760 and villages. *Parasitology* **110 (Pt 2),** 221–229 (1995).
761 48. Criscione, C. D. *et al.* Landscape genetics reveals focal transmission of a human
762 macroparasite. *PLoS Negl. Trop. Dis.* **4,** e665 (2010).
763 49. Betson, M. *et al.* Genetic diversity of Ascaris in southwestern Uganda. *Trans. R. Soc. Trop.*
764 *Med. Hyg.* **106,** 75–83 (2012).
765 50. Betson, M. *et al.* A molecular epidemiological investigation of Ascaris on Unguja, Zanzibar
766 using isoenyzme analysis, DNA barcoding and microsatellite DNA profiling. *Trans. R. Soc.*
767 *Trop. Med. Hyg.* **105,** 370–379 (2011).
768 51. Nejsum, P., Frydenberg, J., Roepstorff, A. & Parker, E. D. Population structure in Ascaris
769 suum (Nematoda) among domestic swine in Denmark as measured by whole genome DNA
770 fingerprinting. *Hereditas* **142,** 7–14 (2005).
771 52. Vercruysse, J. *et al.* Assessment of the anthelmintic efficacy of albendazole in school
772 children in seven countries where soil-transmitted helminths are endemic. *PLoS Negl. Trop.*
773 *Dis.* **5,** e948 (2011).
774 53. Levecke, B. *et al.* The optimal timing of post-treatment sampling for the assessment of
775 anthelminthic drug efficacy against Ascaris infections in humans. *Int. J. Parasitol. Drugs*
776 *Drug Resist.* **8,** 67–69 (2018).
777 54. Zuccherato, L. W., Furtado, L. F., Medeiros, C. da S., Pinheiro, C. da S. & Rabelo, É. M.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

PCR-RFLP screening of polymorphisms associated with benzimidazole resistance in Necator americanus and Ascaris lumbricoides from different geographical regions in Brazil. *PLoS Negl. Trop. Dis.* **12,** e0006766 (2018).

55. Diawara, A. *et al.* Assays to detect beta-tubulin codon 200 polymorphism in Trichuris trichiura and Ascaris lumbricoides. *PLoS Negl. Trop. Dis.* **3,** e397 (2009).

56. Furtado, L. F. V., de Paiva Bello, A. C. P. & Rabelo, É. M. L. Benzimidazole resistance in helminths: From problem to diagnosis. *Acta Trop.* **162,** 95–102 (2016).

57. Olsen, A., Namwanje, H., Nejsum, P., Roepstorff, A. & Thamsborg, S. M. Albendazole and mebendazole have low efficacy against Trichuristrichiura in school-age children in Kabale District, Uganda. *Trans. R. Soc. Trop. Med. Hyg.* **103,** 443–446 (2009).

58. Jaeger, L. H. & Carvalho-Costa, F. A. Status of benzimidazole resistance in intestinal nematode populations of livestock in Brazil: a systematic review. *BMC Vet Res* **13,** 358 (2017).

59. Kaplan, R. M. & Vidyashankar, A. N. An inconvenient truth: global worming and anthelmintic resistance. *Vet. Parasitol.* **186,** 70–78 (2012).

60. Wolstenholme, A. J., Fairweather, I., Prichard, R., von Samson-Himmelstjerna, G. & Sangster, N. C. Drug resistance in veterinary helminths. *Trends Parasitol.* **20,** 469–476 (2004).

61. 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).

62. Easton, A. V. *et al.* Sources of variability in the measurement of Ascaris lumbricoides infection intensity by Kato-Katz and qPCR. *Parasit. Vectors* **10,** 256 (2017).

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

64. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

804 and genome assembly improvement. *PLoS One* **9,** e112963 (2014).

805 65. Otto, T. D., Dillon, G. P., Degrave, W. S. & Berriman, M. RATT: rapid annotation transfer
806 tool. *Nucleic Acids Res.* **39,** e57 (2011).

807 66. Karim, S., Singh, P. & Ribeiro, J. M. C. A deep insight into the sialotranscriptome of the gulf
808 coast tick, Amblyomma maculatum. *PLoS One* **6,** e28525 (2011).

809 67. Duckert, P., Brunak, S. & Blom, N. Prediction of proprotein convertase cleavage sites.
810 *Protein Eng Des Sel* **17,** 107–112 (2004).

811 68. Julenius, K., Mølgaard, A., Gupta, R. & Brunak, S. Prediction, conservation analysis, and
812 structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology* **15,**
813 153–164 (2005).

814 69. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting
815 transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6,** 175–
816 182 (1998).

817 70. Goldberg, J. M. *et al.* Kinannote, a computer program to identify and classify members of the
818 eukaryotic protein kinase superfamily. *Bioinformatics* **29,** 2387–2394 (2013).

819 71. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics*
820 **30,** 1236–1240 (2014).

821 72. Dang, H. X. & Lawrence, C. B. Allerdictor: fast allergen prediction using text classification
822 techniques. *Bioinformatics* **30,** 1120–1128 (2014).

823 73. Saravanan, V. & Lakshmi, P. T. V. Fuzzy logic for personalized healthcare and diagnostics:
824 FuzzyApp--a fuzzy logic based allergen-protein predictor. *OMICS* **18,** 570–581 (2014).

825 74. Dimitrov, I., Bangov, I., Flower, D. R. & Doytchinova, I. AllerTOP v.2--a server for in silico
826 prediction of allergens. *J Mol Model* **20,** 2278 (2014).

827 75. Schwarz, E. M. *et al.* The genome and transcriptome of the zoonotic hookworm Ancylostoma
828 ceylanicum identify infection-specific gene families. *Nat. Genet.* **47,** 416–422 (2015).

829 76. Ghedin, E. *et al.* Draft genome of the filarial nematode parasite Brugia malayi. *Science* **317,**

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1756–1760 (2007).

77. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).

78. Godel, C. *et al.* The genome of the heartworm, Dirofilaria immitis, reveals drug and vaccine targets. *FASEB J.* **26,** 4650–4661 (2012).

79. Desjardins, C. A. *et al.* Genomics of Loa loa, a Wolbachia-free filarial parasite of humans. *Nat. Genet.* **45,** 495–500 (2013).

80. Tallon, L. J. *et al.* Single molecule sequencing and genome assembly of a clinical specimen of Loa loa, the causative agent of loiasis. *BMC Genomics* **15,** 788 (2014).

81. Tang, Y. T. *et al.* Genome of the human hookworm Necator americanus. *Nat. Genet.* **46,** 261–269 (2014).

82. Nemetschke, L., Eberhardt, A. G., Viney, M. E. & Streit, A. A genetic map of the animal-parasitic nematode Strongyloides ratti. *Mol. Biochem. Parasitol.* **169,** 124–127 (2010).

83. Hunt, V. L. *et al.* The genomic basis of parasitism in the Strongyloides clade of nematodes. *Nat. Genet.* **48,** 299–307 (2016).

84. Zhu, X.-Q. *et al.* Genetic blueprint of the zoonotic pathogen Toxocara canis. *Nat. Commun.* **6,** 6145 (2015).

85. Korhonen, P. K. *et al.* Phylogenomic and biogeographic reconstruction of the Trichinella complex. *Nat. Commun.* **7,** 10513 (2016).

86. Mitreva, M. *et al.* The draft genome of the parasitic nematode Trichinella spiralis. *Nat. Genet.* **43,** 228–235 (2011).

87. Foth, B. J. *et al.* Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction. *Nat. Genet.* **46,** 693–700 (2014).

88. Small, S. T. *et al.* Population genomics of the filarial nematode parasite Wuchereria bancrofti from mosquitoes. *Mol. Ecol.* **25,** 1465–1477 (2016).

89. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16,** 157 (2015).

90. Bennuru, S. *et al.* Stage-specific proteomic expression patterns of the human filarial parasite Brugia malayi and its endosymbiont Wolbachia. *Proc. Natl. Acad. Sci. USA* **108,** 9649–9654 (2011).

91. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17,** 10 (2011).

92. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

93. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).

94. Narasimhan, V. *et al.* BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32,** 1749–1751 (2016).

95. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

96. Inbar, E. *et al.* Whole genome sequencing of experimental hybrids supports meiosis-like sexual recombination in Leishmania. *PLoS Genet.* **15,** e1008042 (2019).

97. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **76,** 5269–5273 (1979).

98. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).

99. Dunn, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics* **3,** 32–57 (1973).

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

882 101. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome*
883 *Res.* **19,** 1639–1645 (2009).
884 102. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease
885 and population genetic studies. *Nat. Methods* **10,** 5–6 (2013).
886 103. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using
887 dense haplotype data. *PLoS Genet.* **8,** e1002453 (2012).
888 104. Shaik, J. S., Khan, A. & Grigg, M. E. POPSICLE: A Software Suite to Study Population
889 Structure and Ancestral Determinants of Phenotypes using Whole Genome Sequencing Data.
890 *BioRxiv* (2018). doi:10.1101/338210
891 105. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
892 phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
893 106. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to
894 single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
895 107. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biol.*
896 **13,** R56 (2012).
897 108. Rozas, J. *et al.* Dnasp 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol.*
898 *Evol.* **34,** 3299–3302 (2017).
899 109. Leigh, J. W. & Bryant, D. popart : full-feature software for haplotype network construction.
900 *Methods Ecol. Evol.* **6,** 1110–1116 (2015).
901
902
903
904

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. *Ascaris* germline genome assemblies

Features	*A. lumbricoides de novo*	*A. lumbricoides* semi-*de novo*[1]	*A. lumbricoides* reference-based	*A. suum*[2] (Wang et al 2017)	*A. suum*[2] (Jex et al 2011)[4]
Assembled bases (Mb)	269.2	307.9	296.0	298.0	272.8
N50 (Mb)	0.29	4.77	4.63	4.65	0.41
N50 number	269	21	21	21	179
N90 (Mb)	0.04	0.95	0.91	0.92	0.08
N90 number	1,112	74	75	75	748
Total scaffold number	8,111	412	415	415	29,831
Largest scaffold length (Mb)	1.9	13.9	13.2	13.4	3.8
Protein-coding genes	17,011[3]	17,105[3]	17902	18,025	18,542[3]

[1] Exhibits ~23 Mb of sequence gaps and 15.4 Mb of unplaced sequence in 4,072 short contigs.

[2] The three *A. lumbricoides* assemblies constructed here are compared to the *A. suum* assemblies from Australia [25] and the United States [28].

[3] 21-23% are only partial genes based on the annotation from *A. suum* [28].

[4] The sample for sequencing is derived from a mixture of the germline and somatic genomes (after DNA elimination).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

914 **Table 2. Effects of host, household, village and time point on the genetic variation of *Ascaris***

	Nuclear genome phylogeny[1]			**Mitochondrial genome phylogeny**		
	R^2	**p-value**	**p-adjusted (Bonferroni)**	**R^2**	**p-value**	**Samples[2]**
Individual	0.933	*0.001*	*0.004*	0.996	0.095	68 worms from 60 people
Household	0.020	0.110	0.440	0.011	0.340	68 worms from 43 houses
Village	0.052	*0.001*	*0.004*	0.013	0.335	5 villages with 43, 17, 4, 3, and 1 individual each
Time point	0.018	0.162	0.648	0.024	0.100	55 at baseline and 13 post-deworming

915 [1] Results based on PERMANOVA using phylogenetic distances among worms. Results were largely similar
916 using a distance matrix generated from the PCA plot (Figure S8) and using the Multi-Response Permutation
917 Procedure (MRPP) method (Table S9).

918 [2] Since some worms did not have metadata associated with each variable examined, and some variables were
919 over-represented in the sample (for example, 43 of 68 worms came from a single village) the samples are
920 specified in this column.

921

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure Legends

Figure 1. *Ascaris* proteome

A) Functional classification of the predicted proteome of *A. lumbricoides* (an improved proteome of *Ascaris* spp.), excluding proteins with unknown or uncharacterized function. B) PCA plot based on multivariate analyses of RNA-seq data from various stages/tissues. Samples from tissues related to sperm (blue ellipse) and oocyte production (orange ellipse, see also Figure S9) cluster together. C) Estimated tree based on orthology analyses between the predicted proteomes of publicly available nematodes. The *Ascaris* clade has been shaded in purple within Clade III (teal). Samples are labelled by BioProject Accession number, as well as by the first letter of the genus and the first two letters of the species name (ASU = *Ascaris suum*, ALU = *Ascaris lumbricoides*, WBA = *Wuchereria bancrofti*, BMA = *Brugia malayi*, LLO = *Loa loa*, DIM = *Dirofilaria immitis*, OVO = *Onchocerca volvulus*, TCAN = *Toxocara canis*, ACAN = *Ancylostoma caninum*, ADU = *Ancylostoma duodenale*, ACE = *Ancylostoma ceylanicum*, NAM = *Necator americanus*, CEL = *Caenorhabditis elegans*, SST = *Strongyloides stercoralis*, SRA = *Strongyloides ratti*, TSP = *Trichinella spiralis*, TTR = *Trichuris trichiura*). Multiple genomes for the same organism are suffixed with numerals.

Figure 2. Phylogenetics of *Ascaris* spp based on mitochondrial sequences. A) Haplotype network based on the COl mitochondrial gene. Notches on the lines separating samples represent the number of nucleotide changes between the worms represented, details on the origins of haplotypes can be found in supplementary table S4; B) Maximum likelihood phylogenetic (ML) reconstruction of *Ascaris* complete mitochondrial genomes, constructed under the conditions of the GTR model and 1000 bootstrap replicates were used to provide nodal supports. The tree was constructed using all mitochondrial genomes assembled from the Kenyan isolates and all other published reference *Ascaris* mitochondrial genomes and *Baylisascaris procyonis* was used as the outgroup. The three major clades A, B and C were identified by color hue, and the majority of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

947 Kenyan worms clustered in clade A. Each village was represented by a distinct shape and unfilled
948 shapes represented worms sequenced from specific villages post anthelminthic treatment.

949 **Figure 3: Genetic diversity of the *Ascaris* strains**. A) Circos plot depicting the genetic diversity
950 of the Ascaris strains. Outside track (red histograms) shows the total SNP diversity across the
951 genome (first 50 largest scaffolds) in 10kb sliding windows. Blue bar plot indicates the measured
952 degree of polymorphism (π) [97] within the Ascaris population in 10 kb sliding windows. The
953 innermost track with black-green histogram plots the TajimaD [98] values which reflect the difference
954 between the mean number of pairwise differences (π) and the number of segregating sites using a
955 sliding window of 10 kb. B) The Circos-plot of the genome-wide distribution of heterozygous and
956 homozygous SNPs in 10 kb blocks identified long stretches of homozygosity among the different
957 strains of *Ascaris*, except 119_3, which is predominantly heterozygous throughout and was isolated
958 from village 3. Red color = >90% of heterozygous SNPs, blue = > 90% of homozygous SNPs,
959 yellow = 50% heterozygous, 50% homozygous SNPs. Each track represents a single strain.

960 **Figure 4: Comparative genomics and population genetic structure of *Ascaris***. A) Hierarchy
961 phylogenetic tree of *Ascaris* strains. Phylogenetic tree was constructed with genome wide SNPs (at
962 10x coverage) from 68 *Ascaris* strains, including the *A. suum* reference (outgroup). Height =
963 number of SNPs per site. Red symbol = *A. lumbricoides* mitochondrion genome. Black symbol =
964 *A. suum* mitochondrion genome. Samples were collected from 5 different villages: Circle = village
965 1, square = village 2, upside triangle = village 3, downside triangle = village 4, diamond = village
966 5. **B)** Heatmap clustering the co-inheritance of ancestral blocks by Bayesian method using genome
967 wide shared haplotype segments among the *Ascaris* genomes. scale = posterior coincidence
968 probability. Hierarchical clustering and phylogenetic relationships are based on percent shared
969 haplotype segments in scaffolds ALgV5B01, ALgV5B02, and ALgV5R001. Red arrows show
970 examples of genetic recombination demonstrated by phylogenetic incongruence in the tree
971 topology based on shared ancestry among blue highlighted strains (n = 13). **C)** Pairwise SNPs and
972 F_{ST} estimates in scaffolds ALgV5B01, ALgV5B02, and ALgV5R001 indicate a switching of

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

haplotypes (black arrow), and genetic hybridization among the blue highlighted strains (n = 13) in the phylogenetic tree depicted in figure "B". X-axis = total SNPs/10kb in SNPs plot or F_{ST} /10kb in F_{ST} plot. **D)** Estimation of the number of ancestral populations (K) based on Dunn Index [99]. **E)** Population genetic structure and admixture clustering analysis of the *Ascaris* genomes obtained by POPSICLE [102] using K=6 different color hues in the innermost concentric circle of the Circos plot . The middle concentric circle shows the relative percentage of each genetic ancestry within each genome (represented by the color hues for K = 6). The outermost concentric circle shows the genome wide local admixture profile of each strain in 10 kb sliding windows. The following geometric shapes represent villages, and the color for each shape identifies the mitochondrion genome each sample possesses: Black = *A. suum*; red = *A. lumbricoides*; Circle = village 1; square = village 2; upside triangle = village 3; downside triangle = village 4; diamond = village 5

Figure 5: Local admixture clustering and genome wide analysis of inheritance of haploblocks of *Ascaris* obtained by POPSICLE [104]. Based on ancestral population K= 6. X-axis = strains. Red highlighted box indicates the introgression of large haplotype blocks of defined parentage among the different strains of *Ascaris* in scaffolds ALgV5R019X (A) and ALgV5R027X (B). Many examples exist whereby strains that are in linkage disequilibrium at ALgV5R019X possess different haplotypes in ALgV5R027X (for example 1107E_1 vs. 2110F_2) indicating both segregation as well as recombination in the evolution of the isolates. The local admixture patterns reveal extensive genetic hybridization among different strains of *Ascaris*. Color assignment is depicted based on Figure 6E.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplemental Figure Legends:

Figure S1. Predicted proteome and stage-specific transcriptomes of Ascaris.

A) Functional classification of the predicted proteome of *A. lumbricoides* (an improved proteome of *Ascaris* spp.) with the majority of proteins being unknown/uncharacterized. B) 2-dimensional principal component analysis plot illustrating the similarities in transcription profiles between the major stages (Fig 1B) and the developmental stages.

Figure S2. Phylogenetic trees based on CO1 and ND4

Maximum likelihood phylogenetic analyses of the A) CO1 and B) ND4 genes using RaxML under the conditions of the GTR model with nodal support values generated through 1000 bootstrap replicates. The trees were generated using complete sequences of the genes extracted from the 68 Kenyan *Ascaris* mitochondrial genomes generated in this study and other published reference genomes.

Figure S3. Sliding window analyses. A) Comparison between Kenyan samples and reference mitochondrial genomes of *Ascaris lumbricoides* and *Ascaris suum*, B) Comparison between villages, C) CO1 comparison between villages.

Figure S4. *Ascaris* SNPs and insertion/deletions (InDels) maps.

An assembled 6.5 Mb *Ascaris lumbricoides* chromosome fragment (ALgV5R006), with the frequency of identified SNPs and InDels plotted for one representative *A. lumbricoides-like* worm from this study (#7664) and one *A. suum-like* worm (#7680). Genes are shown on the top of the plot, with red and blue indicating genes transcribed from forward and reverse strands, respectively. The y-axis shows the frequency of SNPs and InDels for a 20-kb window size (with a 4-kb sliding window in x-axis). Note the profiles and the frequency between SNPs and InDels are highly consistent within individual worms.

Figure S5. Somy analysis of the *Ascaris* isolates. The ploidy of the *Ascaris* strains are represented in a heatmap. Ploidy was calculated by averaging the count of aligned reads in 10 kb sliding windows across the genome after reference mapping against ALV5. The ploidy data suggest that

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ascaris is completely diploid (close to 2n), except at two scaffolds ALgB5B14 and ALgv5RO23, where the majority of strains show elevated ploidy. X-axis shows the first 50 largest scaffolds involved in this study and the y-axis shows the strains (ordered by code number 1-68).

Figure S6. Admixture clustering and current population genetic structure of Ascaris were determined. Data analyzed with POPSICLE with an ancestral population size = 4 (A) and 8 (B) in 10 kb sliding windows as described in Figure 4E.

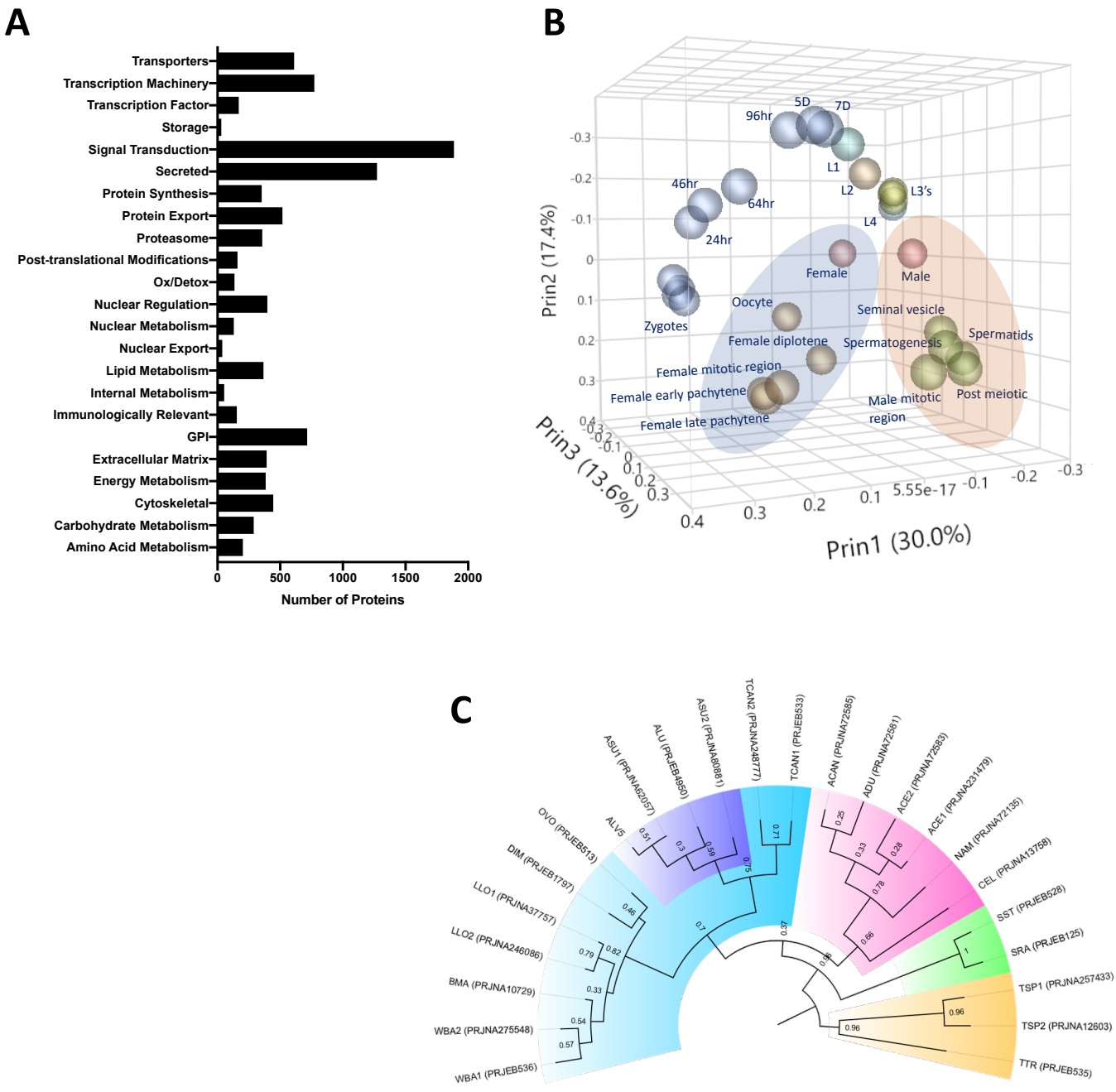
Figure S7. Plot of phylogenetic distances compared to geographic distances. A) For village #1 and village #5. B) Plot of diversity versus geographical distance (Hs on left, Fst on right). Genetic distances based on CO1 genes are plotted against the geographic distances between the places from which these worms were collected.

Figure S8. PCA plot of worms sequenced for 5 Kenyan villages. Each point is color-coded by village-of-origin and plotted according to the 1st and 2nd principal components, based on genome sequences. Worms from village #1 are found in each of 3 clusters, and two clusters contain only worms from village #1.

Figure S9. *Ascaris* stage-specific RNA expression heatmaps (A) Correlation heatmap comparing parasite transcriptomes at different life stages. (B) 1870 genes differentially expressed across the stages.

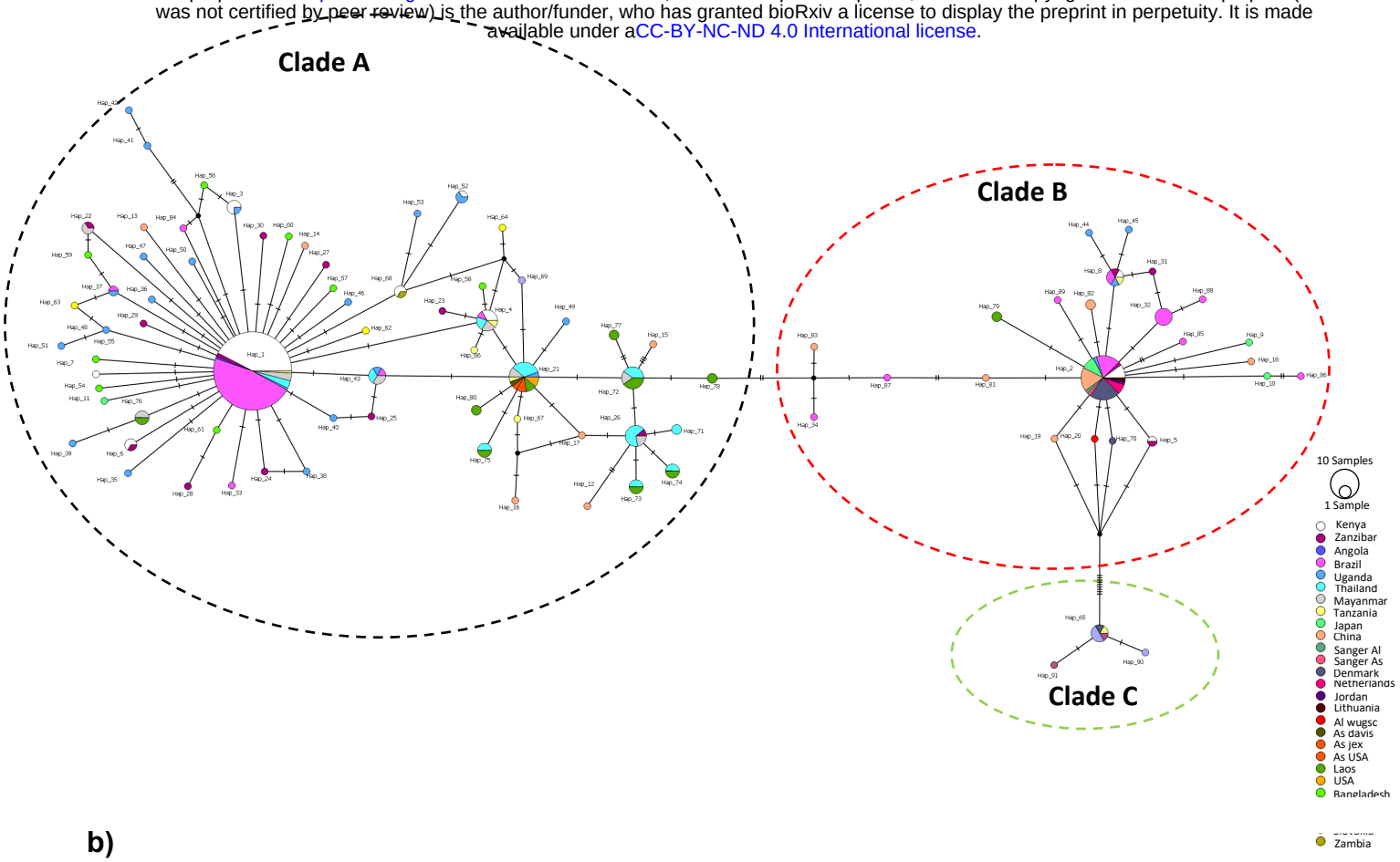
Figure S10. Evidence of *Ascaris* population expansion. The pairwise nucleotide differences between worm samples (solid line) are compared to the binomial function that would most closely represent a theoretical stable population (dotted line). Additional information is available in Table S8.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

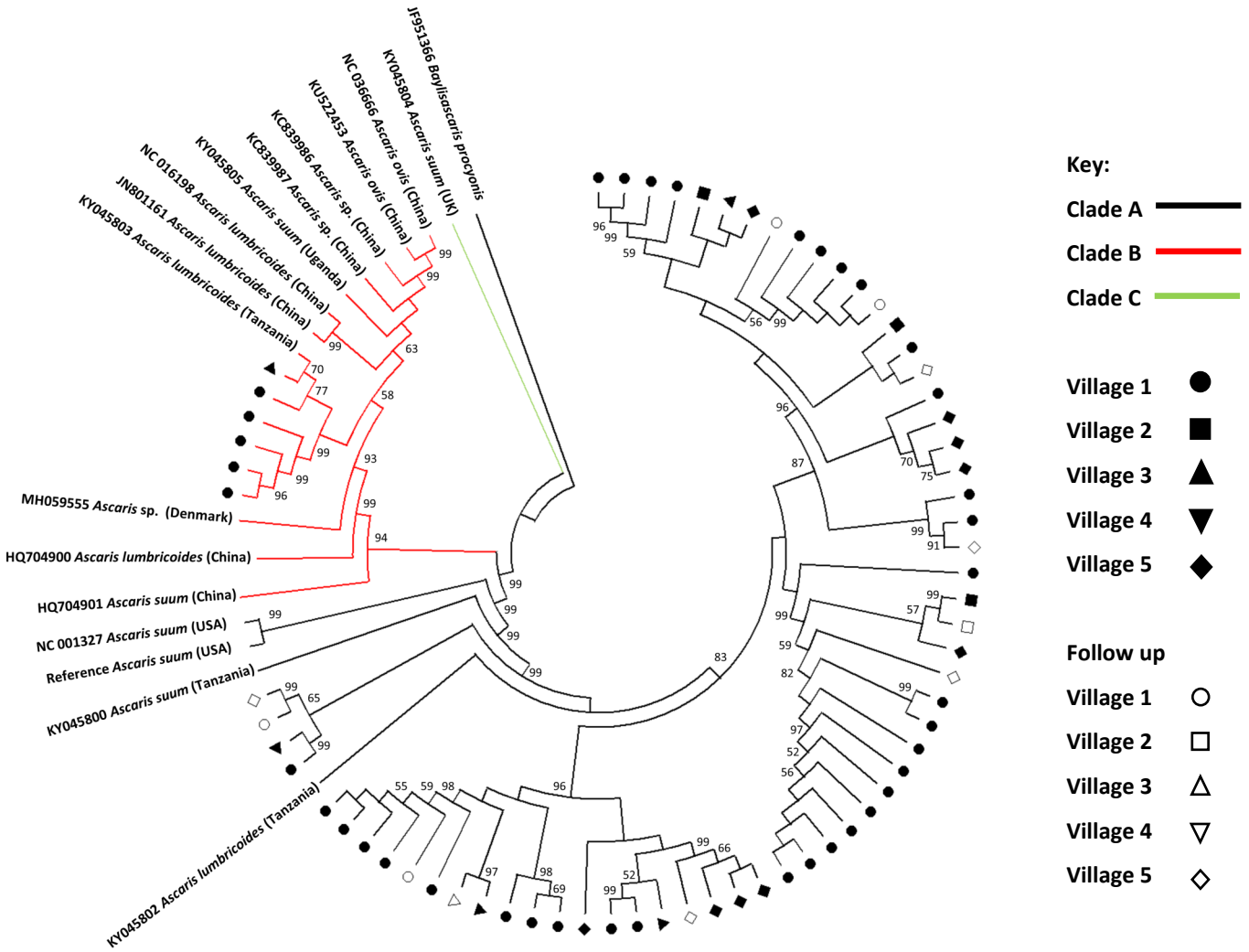


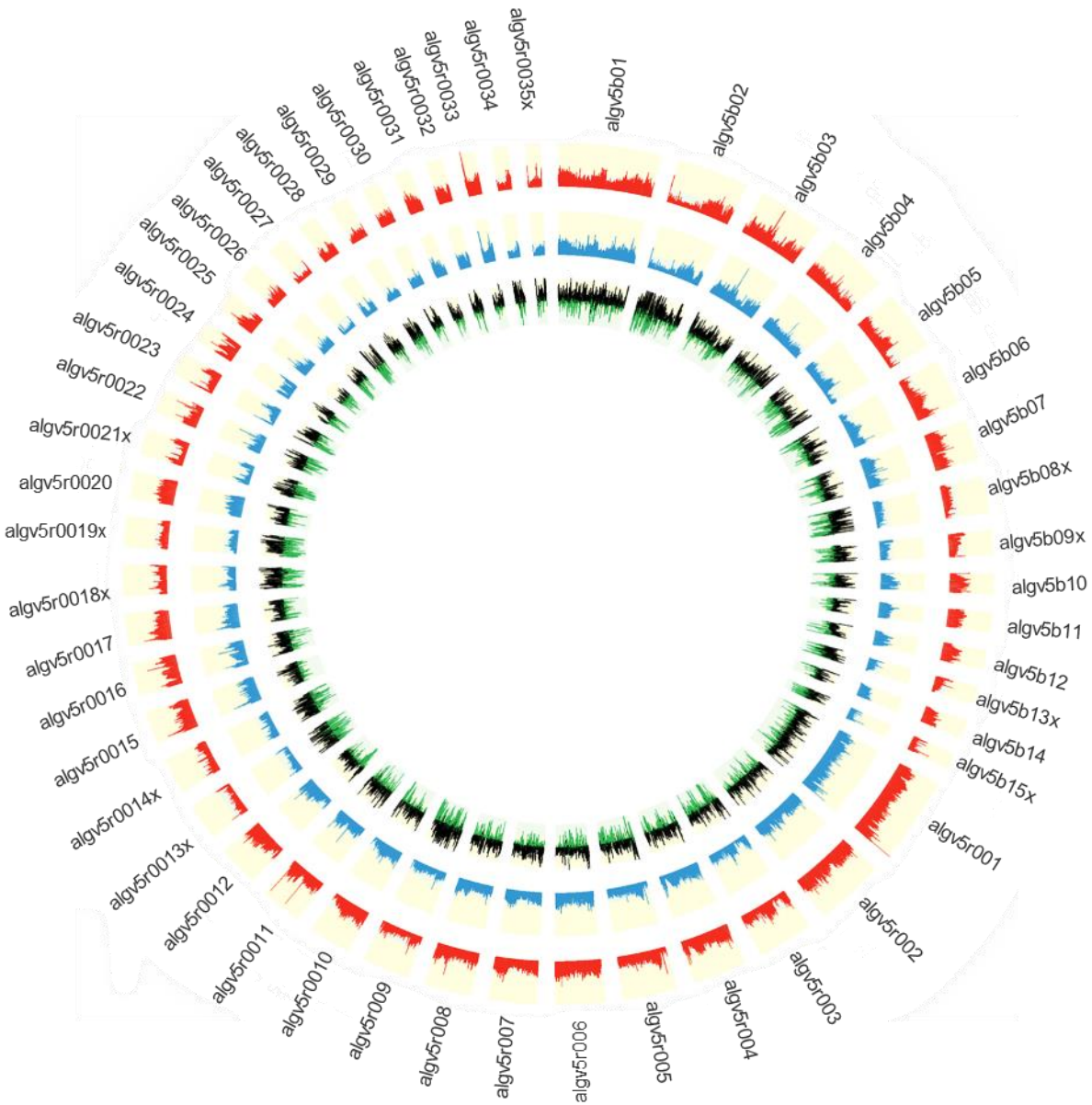
bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

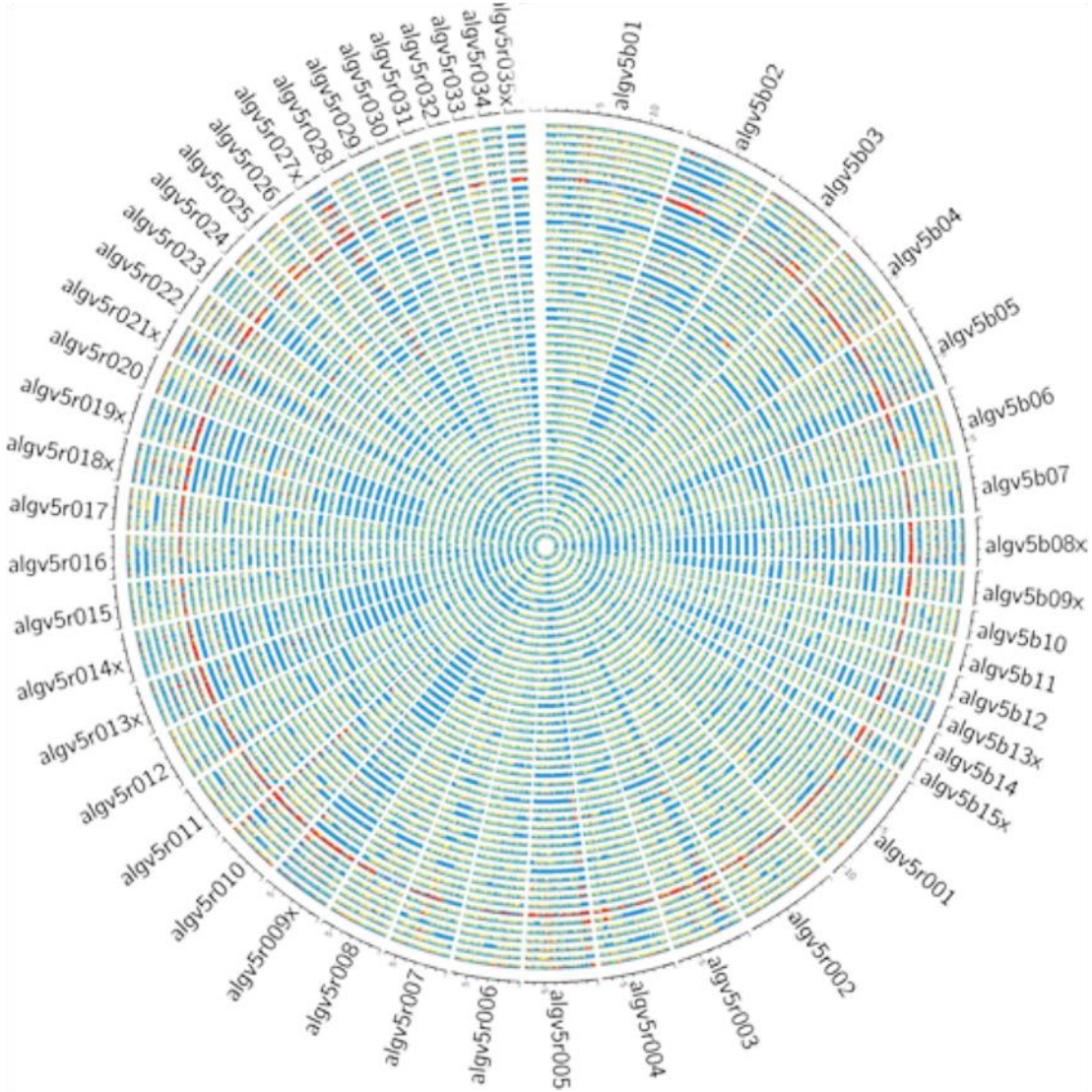
a)



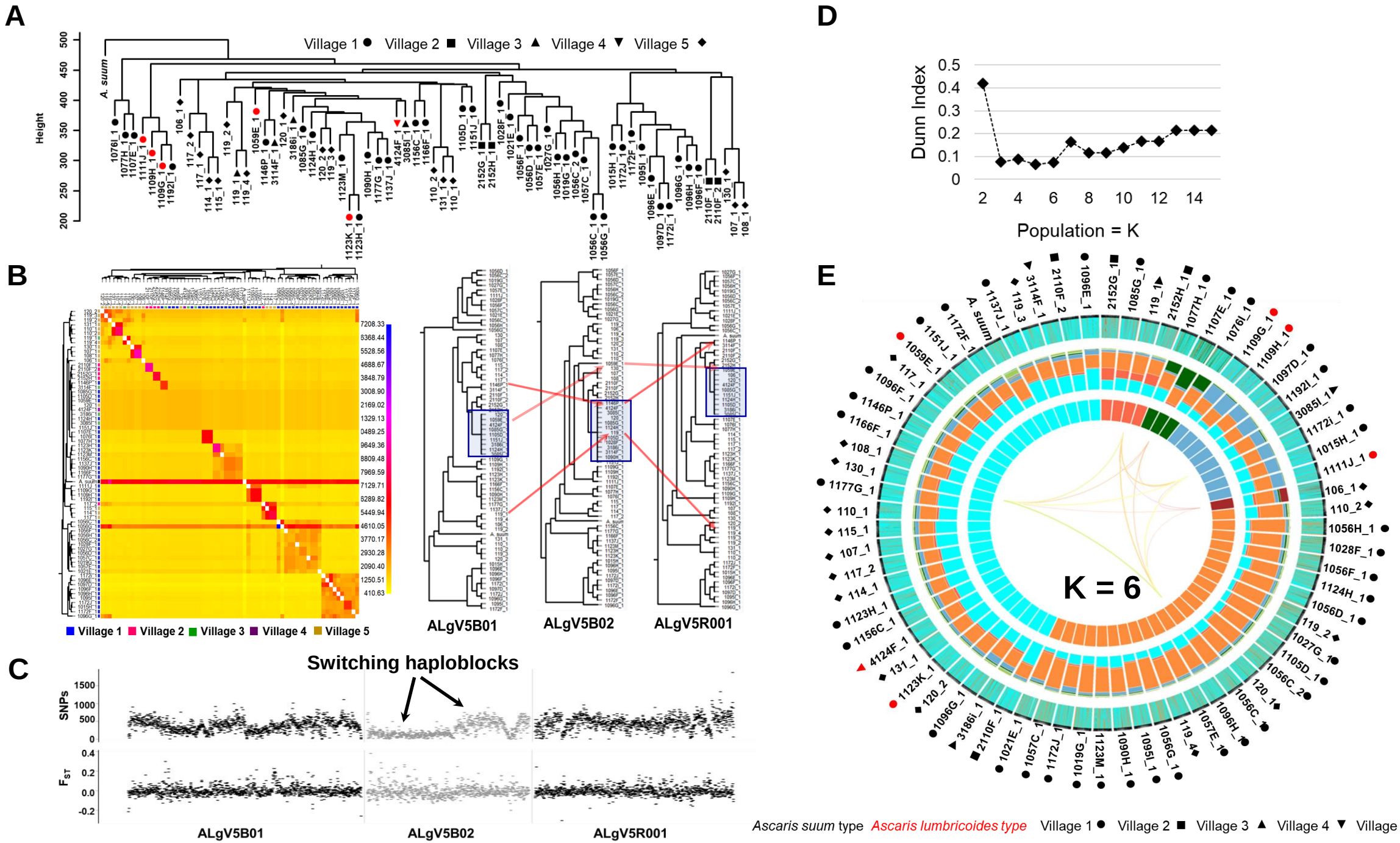
b)



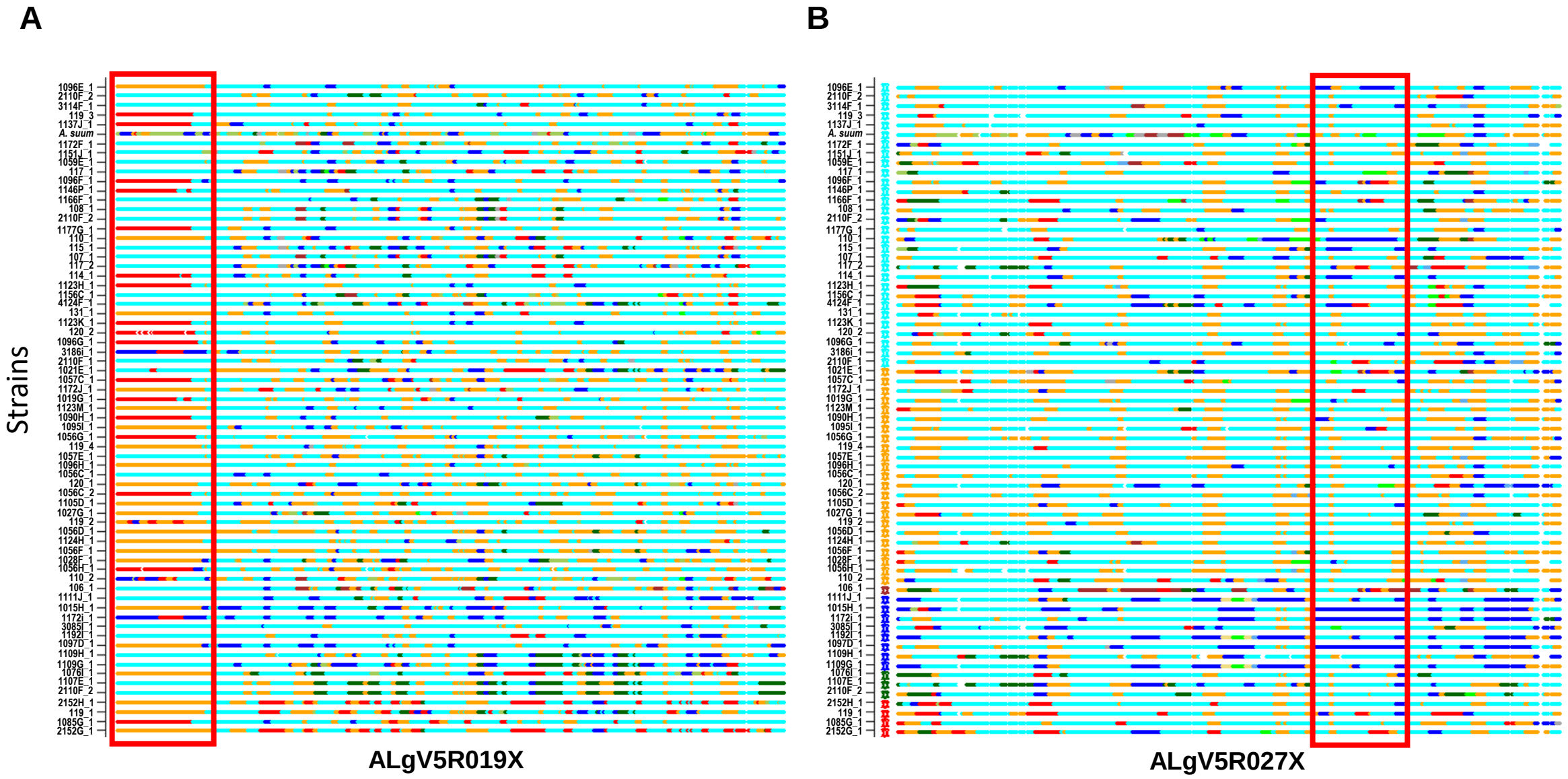

algv5b01
algv5b02
algv5b03
algv5b04
algv5b05
algv5b06
algv5b07
algv5b08x
algv5b09x
algv5b10
algv5b11
algv5b12
algv5b13x
algv5b14
algv5b15x
algv5r001
algv5r002
algv5r003
algv5r004
algv5r005
algv5r006
algv5r007
algv5r008
algv5r009
algv5r0010
algv5r0011
algv5r0012
algv5r0013x
algv5r0014x
algv5r0015
algv5r0016
algv5r0017
algv5r0018x
algv5r0019x
algv5r0020
algv5r0021x
algv5r0022
algv5r0023
algv5r0024
algv5r0025
algv5r0026
algv5r0027
algv5r0028
algv5r0029
algv5r0030
algv5r0031
algv5r0032
algv5r0033
algv5r0034
algv5r0035x


algv5b01
algv5b02
algv5b03
algv5b04
algv5b05
algv5b06
algv5b07
algv5b08x
algv5b09x
algv5b10
algv5b11
algv5b12
algv5b13x
algv5b14
algv5b15x
algv5r001
algv5r002
algv5r003
algv5r004
algv5r005
algv5r006
algv5r007
algv5r008
algv5r009x
algv5r010
algv5r011
algv5r012
algv5r013x
algv5r014x
algv5r015
algv5r016
algv5r017
algv5r018x
algv5r019x
algv5r020
algv5r021x
algv5r022
algv5r023
algv5r024
algv5r025
algv5r026
algv5r027x
algv5r028
algv5r029
algv5r030
algv5r031
algv5r032
algv5r033
algv5r034
algv5r035x

Figure


A
Village 1 Village 2 Village 3 Village 4 Village 5
A. suum
Height
B
Village 1 Village 2 Village 3 Village 4 Village 5
ALgV5B01
ALgV5B02
ALgV5R001
C
Switching haploblocks
SNPs
F_{ST}
ALgV5B01
ALgV5B02
ALgV5R001
D
Dunn Index
Population = K
E
K = 6
Ascaris suum type
Ascaris lumbricoides type
Village 1 Village 2 Village 3 Village 4 Village 5

Supplemental figure



bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary text

Ascaris lumbricoides reference genome

We first evaluated a *de novo*-only assembly strategy. To facilitate scaffolding, we obtained Pacific Biosciences (PacBio) reads from the female germline (ovary and oviduct) and intestinal tissues. Despite the high coverage of the Illumina paired-end and mate-pair reads and additional longer PacBio reads, the *de novo*-only assembly approach led to a low-quality draft genome with a small N50 value and many scaffolds. We next used a *de novo* assembly strategy with scaffolding to the reference *A. suum* genome (semi-*de novo* strategy). While this assembly led to relatively high N50 values, the assembly contained ~23 Mb of sequence gaps and an additional 15.4 Mb of sequence that could not be incorporated into the assembly. We believe the reference-based *ALV5*. assembly describe in the main text is the highest quality assembly among the three versions (Table 1) and it should be used for future studies. However, we note that several caveats should be considered when using this genome assembly: (1). potential assembly errors in the reference *A. suum* genome could be propagated into the *ALV5* genome; (2) potential sequence differences between AlV5 and *A. suum*, particularly large-scale genome rearrangements and structure variations, are not likely incorporated into the *ALV5* assembly; and (3) annotation errors and/or missed annotation, either due to errors in the original *A. suum* annotation or due to the potential different gene expression profiles between the two species may be present in the *ALV5* genome. To identify assembly errors and structure variations, we have used additional Illumina sequencing reads to further improve the genome assembly using Pilon genome polishing software. Pilon uses a local re-assemble strategy to correct draft assembly and to call sequence variants of multiple sizes, including very large insertions and deletions [1]. Overall, the *ALV5* genome assembly is reference-quality and is suitable for downstream analysis, including SNPs and InDels identification. All genome assemblies are made available (see section *Data availability*).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Optical mapping, long sequencing reads, and three-dimension chromosomes organization data (Hi-C) will be necessary to further improve the assembly.

Genome annotation and *Ascaris* proteome

In addition to transferring *A. suum* annotations to the reference-based germline *A. lumbricoides* genome, we also transferred the *A. suum* gene models to the *de novo* and semi-*de novo A. lumbricoides* genome versions. While ~94.6% of the genes can be transferred to both genomes, over 20% of the transferred genes are only partial matches and are fragmented supporting the view that the *de novo* and semi *de novo A. lumbricoides* assemblies are highly fragmented. Although the updated *Ascaris suum* proteome was reported in our previous study [2], that study's focus was on comparative analysis of programmed DNA elimination. Thus, no in-depth analysis of the revised *Ascaris* proteome (based on the annotation of the reference-based genome) was carried out in that study.

Nucleotide diversity and population structure inferred from mitochondrial genomes

Sliding window analyses showed considerable diversity across *Ascaris* mitochondrial genomes. The most diverse region was the non-coding control region, both between worms from different villages in this study and between Kenyan and other worms. The ND4 gene had a nucleotide diversity of $\pi = 0.008827$; that of the CO1 barcoding gene was 0.006243 (Figure S2). Despite differences in their molecular diversity, phylogenetic analyses based on the ND4 and CO1 genes revealed the same overall topology: distinct *A. lumbricoides* and *A. suum* type clades (Figure S3). Though the phylogeny passed on the whole mitochondrial genomes also suggests the presence of an *A. lumbricoides*-type clade and an *A. suum*-type clade, the mitochondrial genome generated from the Sanger institute *A. suum* genome reads (Acc: PRJEB2435) clustered into the *A. lumbricoides* type clade (Figure 3B).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Evolution of *Ascaris* mitochondrial genomes

To account for potential population expansion events, both datasets including all published mitochondrial genomes from across the globe and those from Kenya alone were assessed (Figure S10 and Table S8). The Tajima's D value was negative and significant (Tajima's D -1.5691; P-value 0.028, P<0,05) indicating an excess of low frequency polymorphisms within the global population data set. The Fu's *Fs* for the global data was shown to be positive but not significant (Fu's *Fs* 8.5673; P-value 0.98, P>0.05), potentially indicating a deficiency in diversity as would be expected in populations that has recently undergone a bottle neck event. Furthermore, the same patterns were also seen when only the Kenyan sequences were compared to each other where the Fu's F_S was positive and nonsignificant (Fu's F_S 4.979; P-value 0.917, P>0.05), and the Tajima's D [43] was also negative but nonsignificant (Tajima's D -1.28930; P-value 0.079, P>0.05 .

Overall, these data suggest that *Ascaris* either underwent a population expansion or ~~a~~ significant selective pressure. To test for selection on the *Ascaris* population, we calculated the ratio of non-synonymous to synonymous substitutions (dN/dS ratio) for each of the genes in the mitochondrial genome. All genes appeared to be under purifying selection with a dN/dS ratio <1 (with ND6 having the lowest ratio value of 0.05 and ND5 with the highest ratio value of 0.527). Since we observed a dN/dS ratio below 1 for all genes, it is unlikely that the genetic variability of the *Ascaris* samples examined here resulted from selective sweeps. This means that these genes are likely under purifying selection to maintain the existing functionality of these mitochondrial-encoded proteins, as opposed to drastic changes resulting from selective sweeps. Thus, the Tajima's *D* and Fu's *F* measurements are most likely due to a population expansion. Since we do not know how quickly *Ascaris* mutates, we cannot estimate how recently such a population expansion might have occurred. Results were

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

similar when based on sequences of worms collected only in Kenya, suggesting that a recent (worldwide) *Ascaris* population expansion may have occurred.

Population structure inferred from nuclear genomes

Most clades in the nuclear genome phylogeny are dominated by worms from either village #1 or village #5, which were the most heavily parasitized and also the most heavily represented in this sampling. P-values for clustering of worms from the same host-groups (Table 2) are shown with Bonferroni corrections for multiple comparisons. The two significant groupings (for the nuclear phylogeny) were individual and village. These remain significant if both factors are included in the same regression, or if the "individual" variable is treated as being nested within the "village" variable.

Geographic distance was not a significant predictor of phylogenetic distance across the five villages (r2=0.07, p=0.15) within village #5 (r^2=0.05, p=0.27), or within village #1 (r^2=-0.07, p=0.81). This analysis was only done for these two villages, as they had the largest number of worms sequenced in this study (Figure S6).

Supplementary Methods

DNA extraction from adult worms

DNA extraction methods were modified several times before a protocol was identified that produced DNA of sufficient quality. Extraction was likely difficult because the worms have a high polyphenol content [3]. In the end, approximately 200-400 mg of worm material was put in a 1.5 mL tube (Eppendorf, Hamburg, Germany). To the sample, 750 µl of Qiagen Digestion buffer G2 (Qiagen, Hilden, Germany) and 20 µl of Proteinase K (Qiagen, Hilden, Germany) was added. Tubes were manually shaken for ten seconds, and then incubated in a Thermomixer C heat block (Eppendorf, Hamburg, Germany) at 50 degrees and 500 rpm for

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

four hours. Halfway through the incubation, tubes were manually shaken again. Tubes were spun at 10000 rpm for one minute, and 700 µl of the supernatant was removed and to which 700 µl one-phase Phenol/Chloroform/Isoamyl alcohol (Amresco LLC, Solon, Ohio, USA) was added to these tubes. Tubes were inverted 30 times, then spun 4^0C at 14000 rpm for 20 minutes. The supernatant was removed, and 3 µl RNase A (ThermoFisher Scientific, Waltham, MA, USA) was added. The sample was mixed and incubated at 37^0Cfor 30 minutes at 200 rpm. Once again, 700 µl one-phase Phenol/Chloroform/Isoamyl alcohol was added to these samples, the samples were inverted 30 times, and spun for 15 minutes. The supernatant was then removed and added to the Zymo Genomic DNA Clean & Concentrator kit.

Analysis of mitochondrial genomes

Genetic diversity was compared to the CO1 gene, as the CO1 gene has been used extensively in molecular epidemiological studies of *Ascaris* and other nematodes. The sliding window was performed on all mitochondrial genomes and the subset containing only those from Kenya. For comparison, sliding window analyses were also performed on the complete CO1 gene in order to identify if the barcoding region itself was actually the most diverse region of the gene to be used.

Non-coding and tRNA sequences were removed from the analyses, to remove the potential effect of high mutation rates which could have caused mutational saturation in any of the analyses and masked the true impact of the divergence in the protein coding genes.

Supplementary Tables

Table S1. Characteristics of genome assemblies

Table S2. Proteome annotation

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

126 **Table S3. Description of worm from which each sample was sequenced.** The sex of the
127 worm (based on morphological identification) and the part of the worm (germline vs somatic)
128 is listed. Some hosts donated multiple worms.

129 **Table S4. CO1 haplotype list**

130 **Table S5. X4 ratio analyses of Clade A and B using complete mitochondrial genomes**
131 **used to construct the phylogeny in Fgure 2b.**

132 **Table S6. Number of heterozygous and homozygous SNPs in each of the 68 worms from**
133 **Kenya sequenced.**

134 **Table S7. Reference mitochondrion genomes**

135 **Table S8. Demographic analyses using Tajima's D and Fu's F statistic across complete**
136 **mitochondrial genomes as a detection for the signature of population expansion events.**

137 **Table S9. Supplement to Table 2 using alternative measures of phylogenetic distance**

138

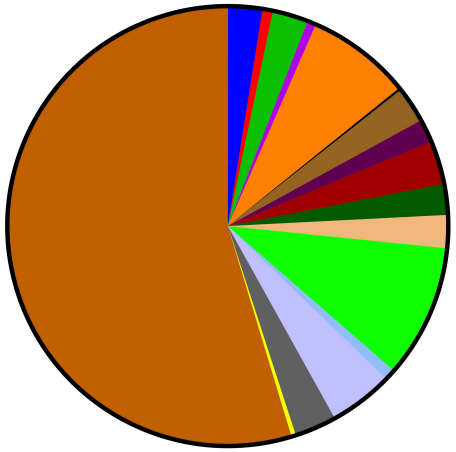
139 **Supplementary figure legends in main manuscript file**

140 **Bibliography**

141 1. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection
142 and genome assembly improvement. *PLoS One* **9,** e112963 (2014).
143 2. Wang, J. *et al.* Comparative genome analysis of programmed DNA elimination in
144 nematodes. *Genome Res.* **27,** 2001–2014 (2017).
145 3. Di Mito, C. & Betschart, B. DNA extraction from Ascaris suum muscle tissue. *Parasitol.*
146 *Res.* **84,** 596–597 (1998).

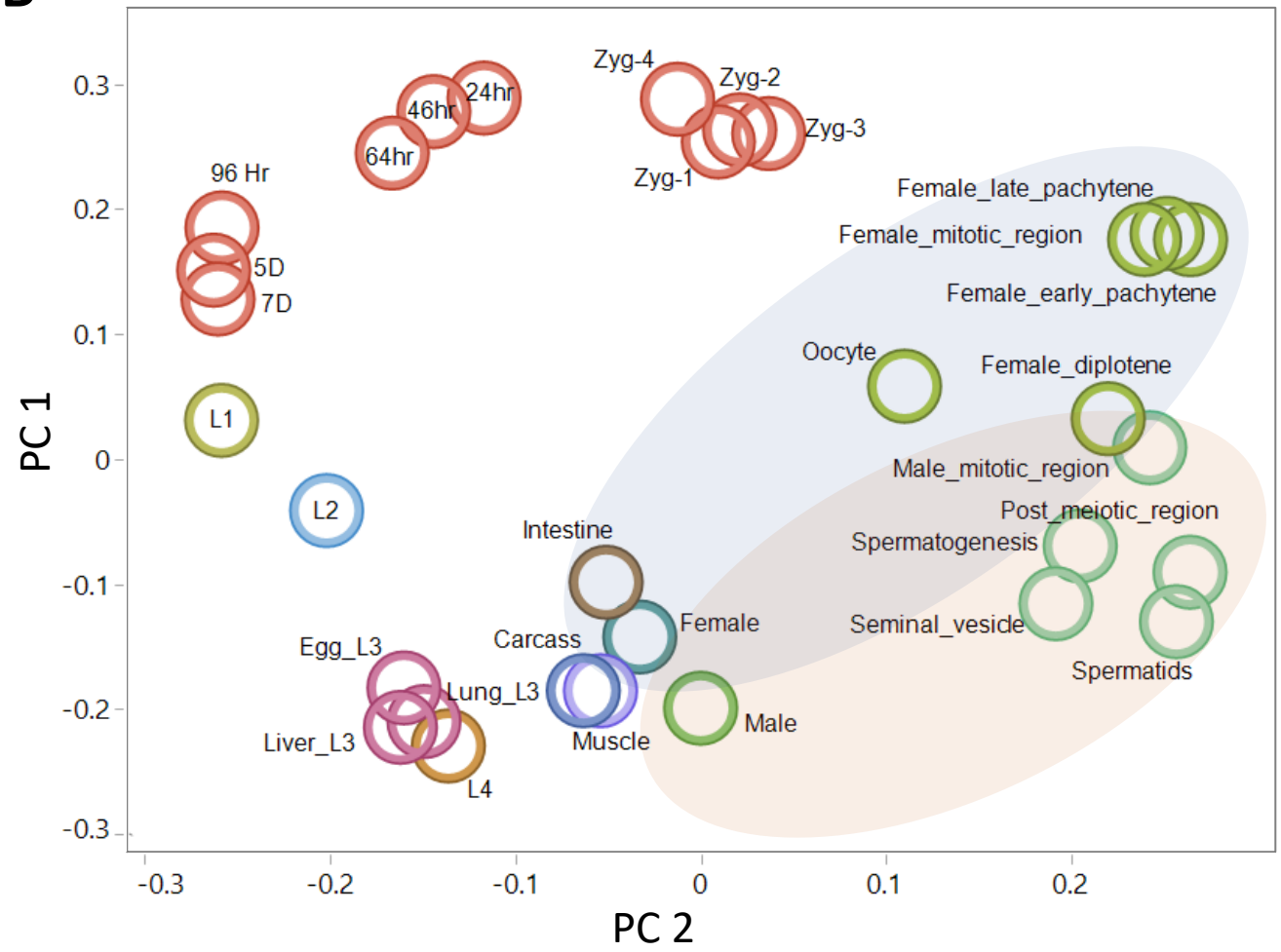
147

A



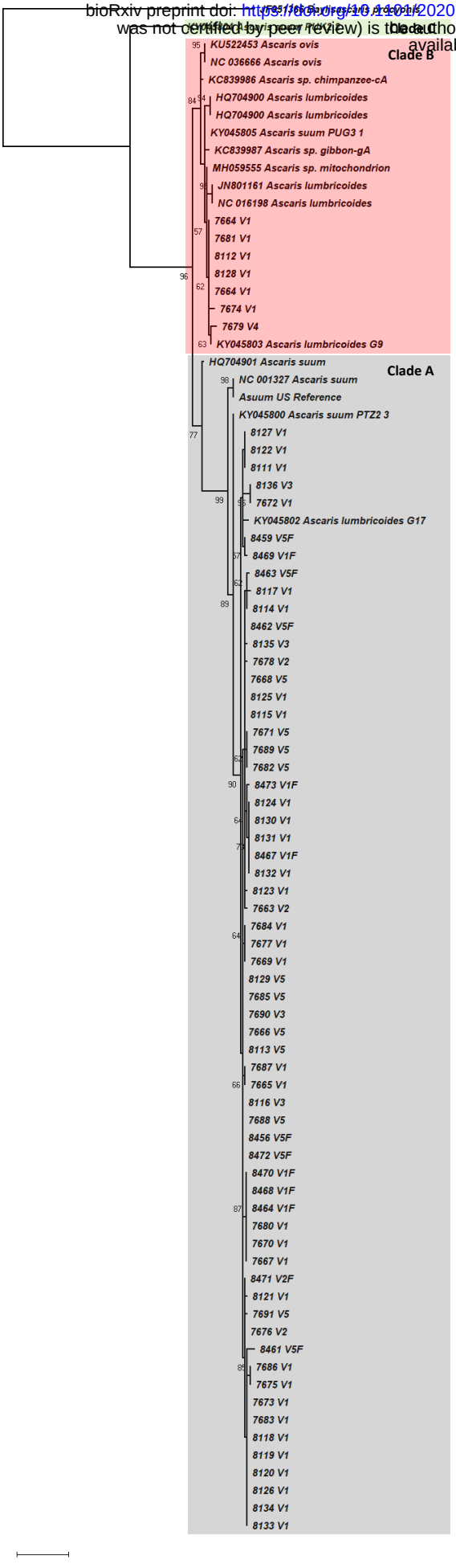
- Cytoskeletal
- Detoxification
- Extracellular Matrix
- Immunologically Relevant
- Metabolism
- Nuclear Export
- Nuclear Regulation
- Proteasome
- Protein Export
- Protein Synthesis
- Secreted
- Signal Transduction
- Transcription Factors
- Transcription Machinery
- Transporters
- Transposable Elements
- Uncharacterized
- Viral

B

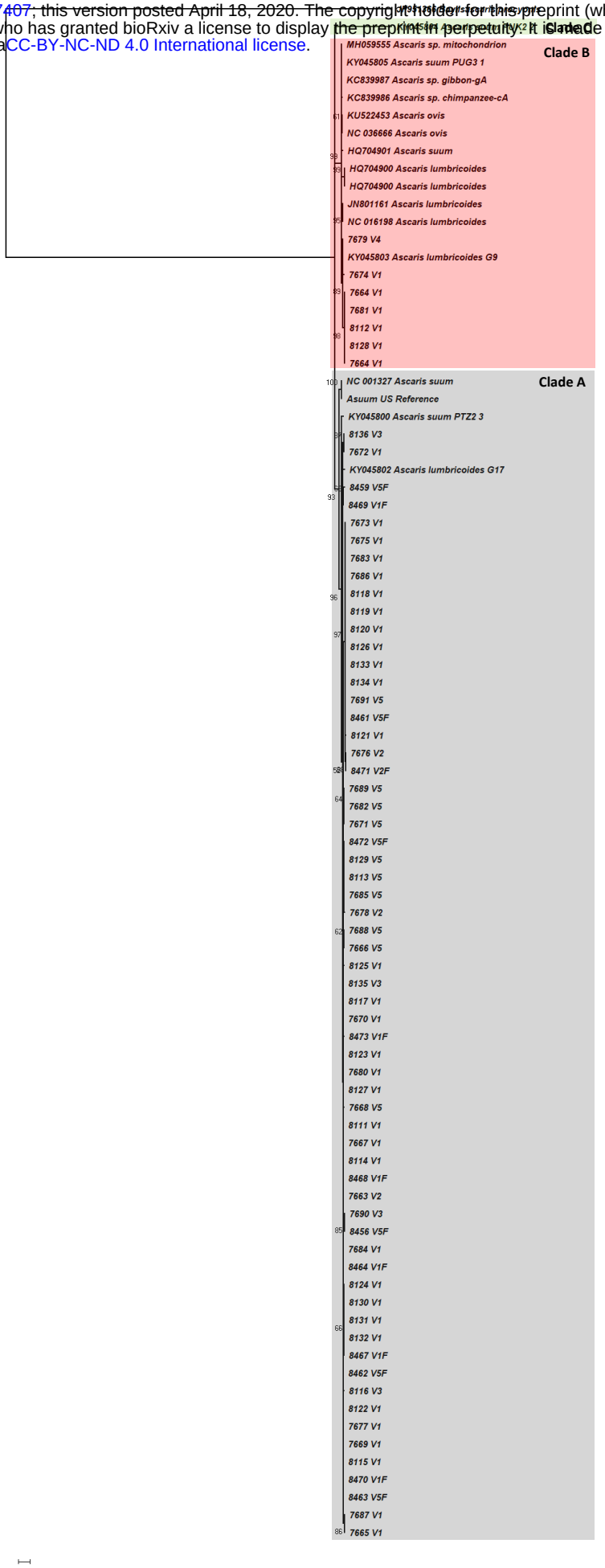


bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

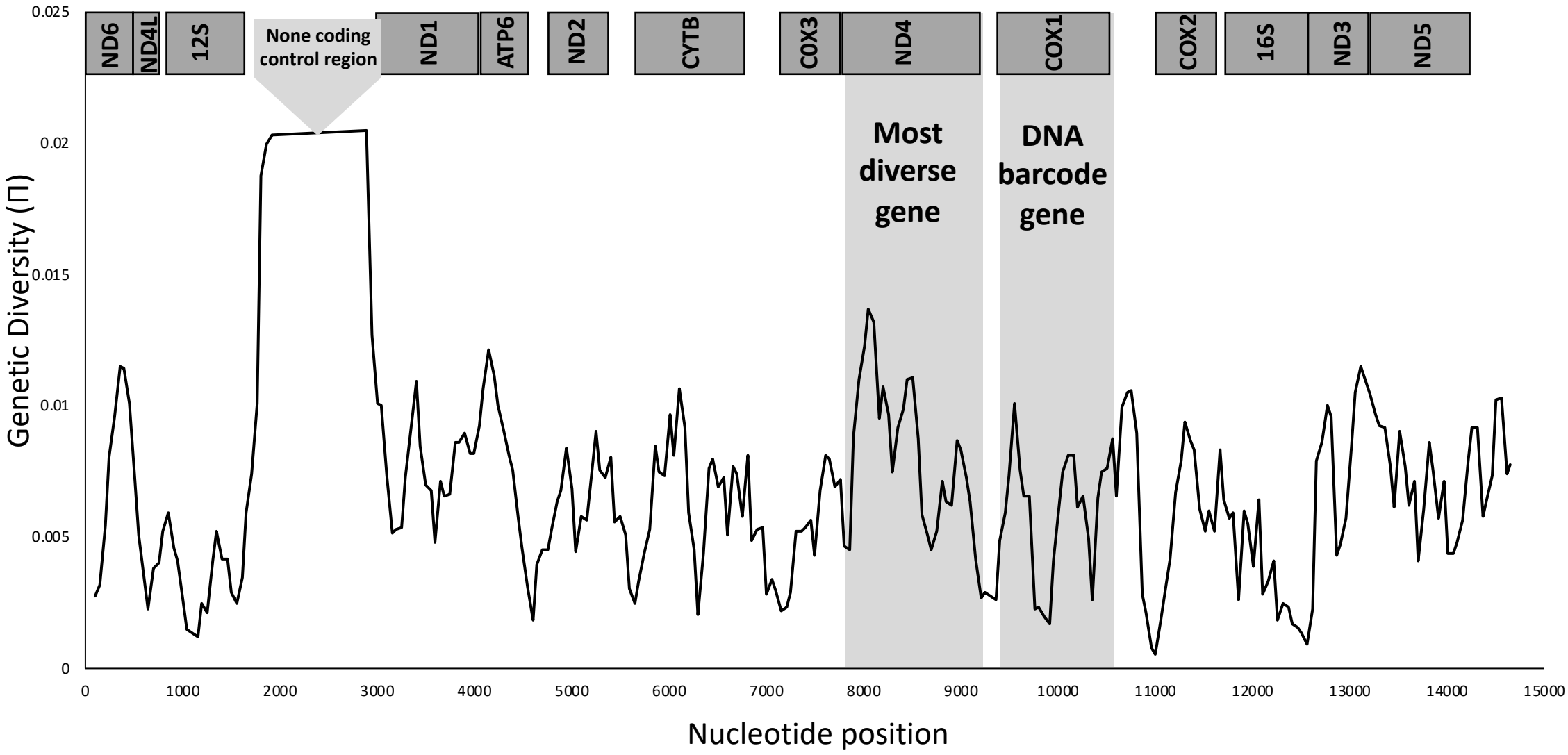
a) CO1



b) ND4

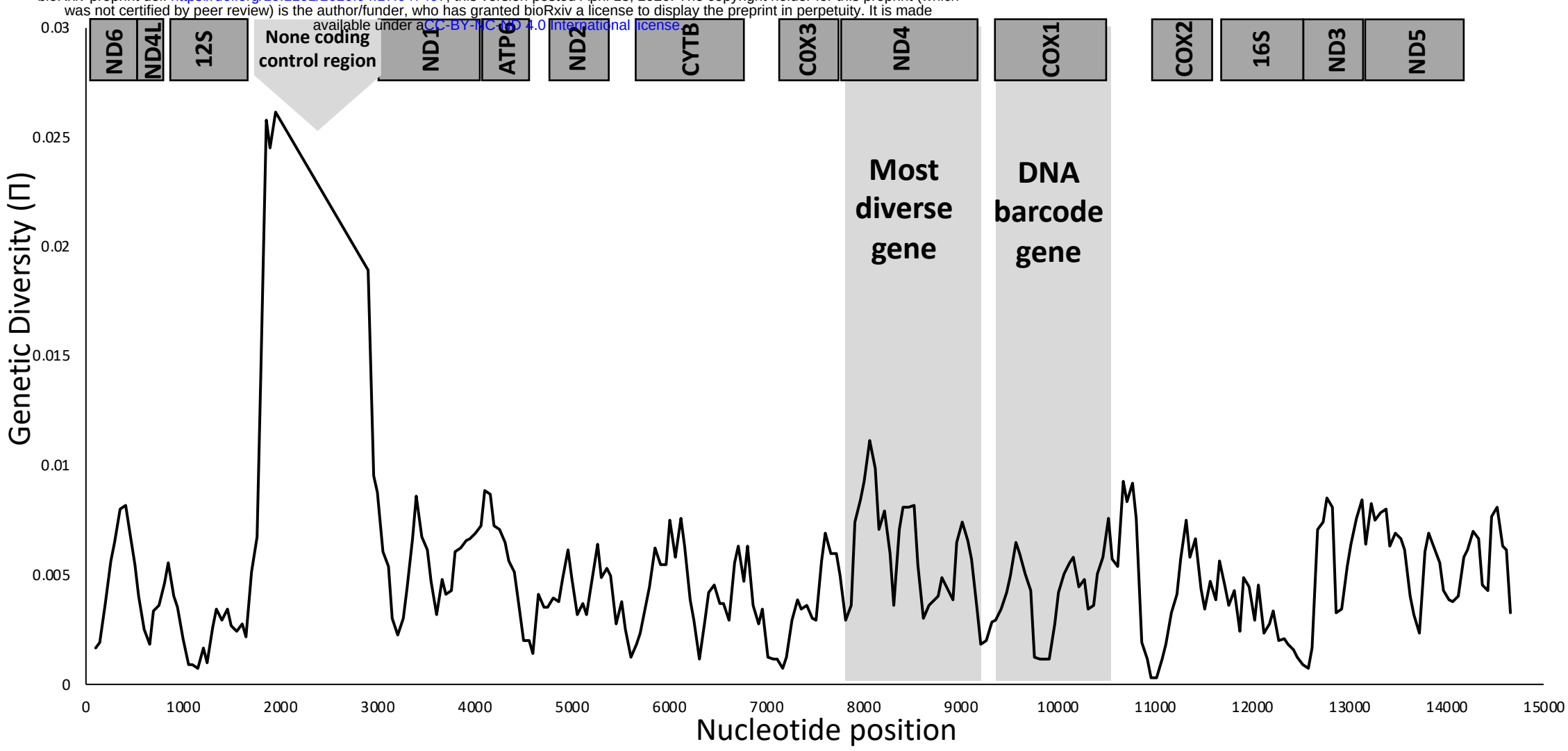


A) Comparison between Village samples and reference mt genomes of *Ascaris lumbricoides* and *Ascaris suum*

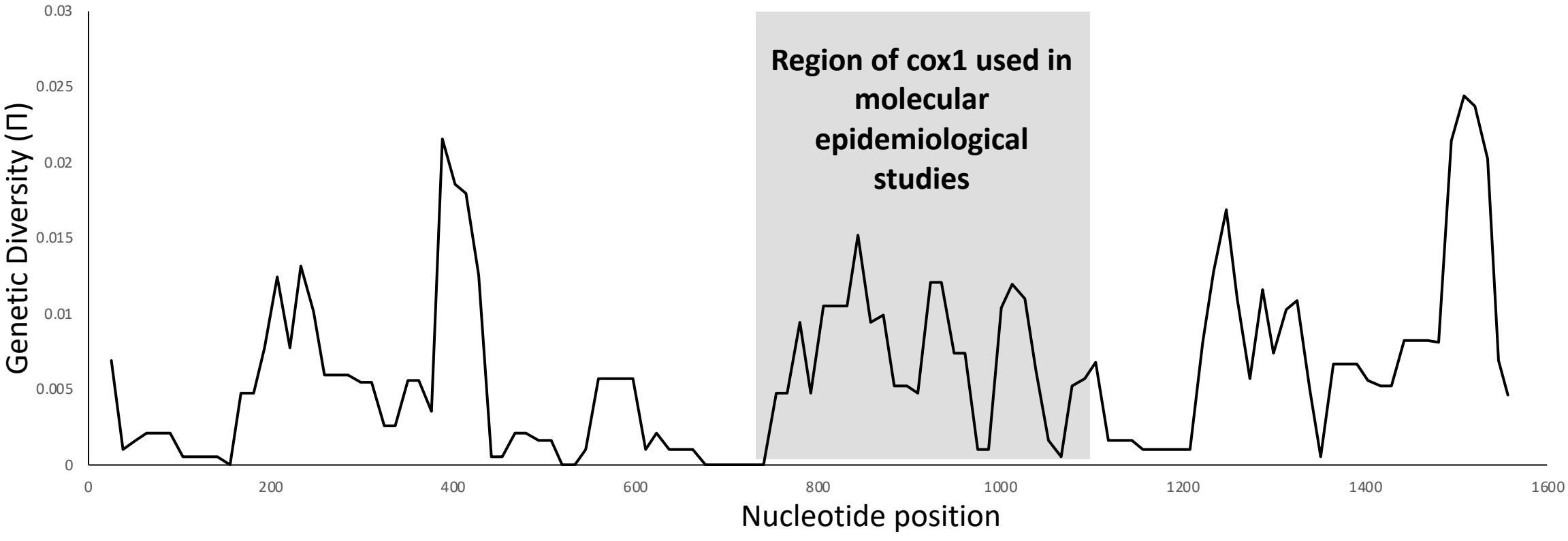


B) Comparison between Villages

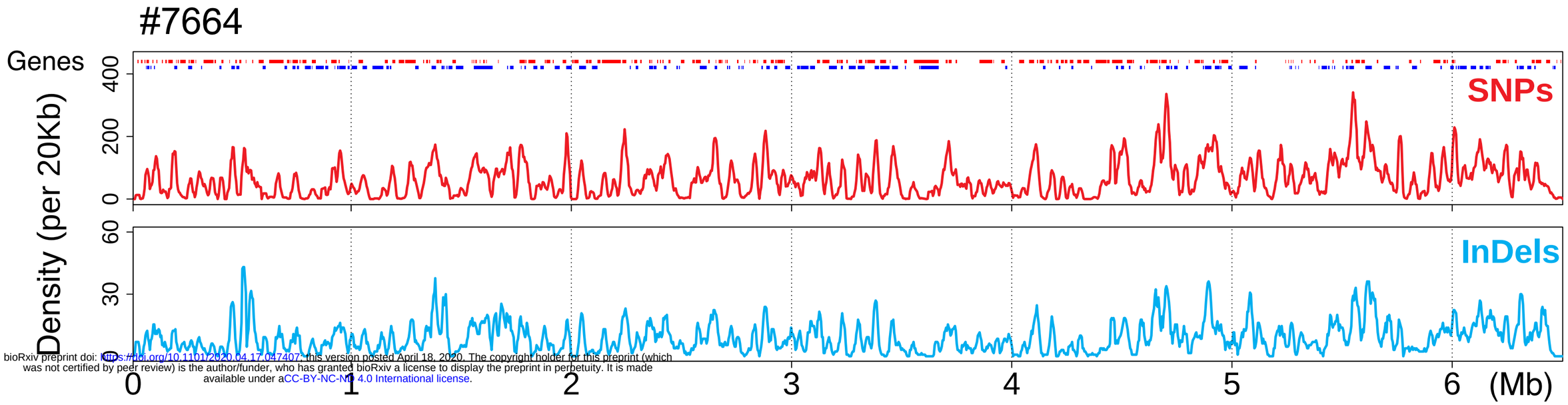
bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

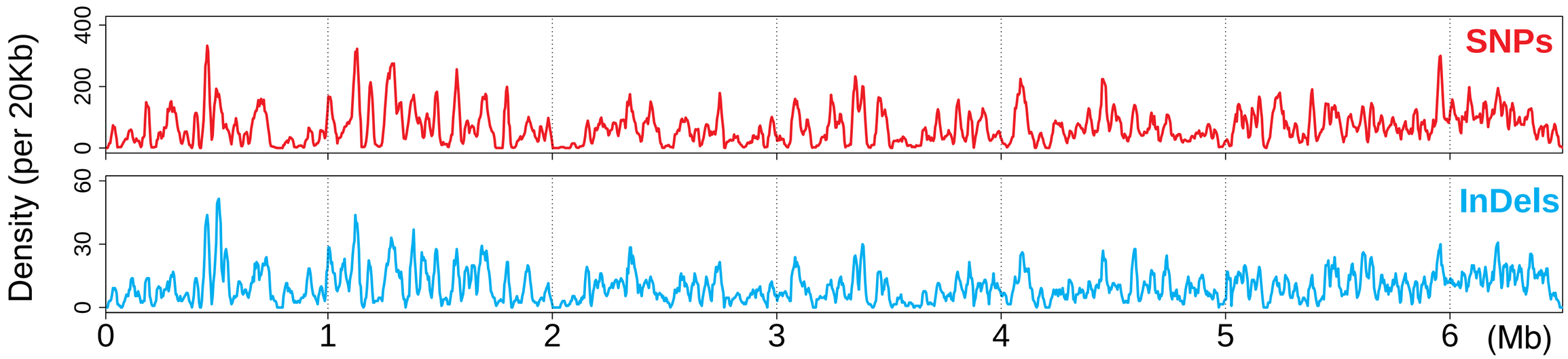


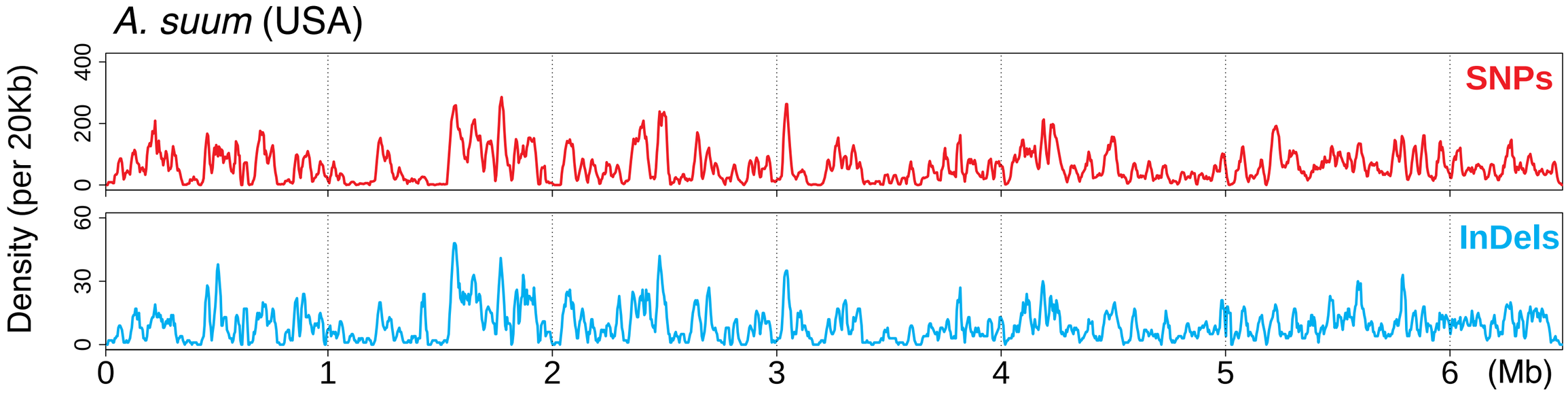
C) Cox1 comparison between villages

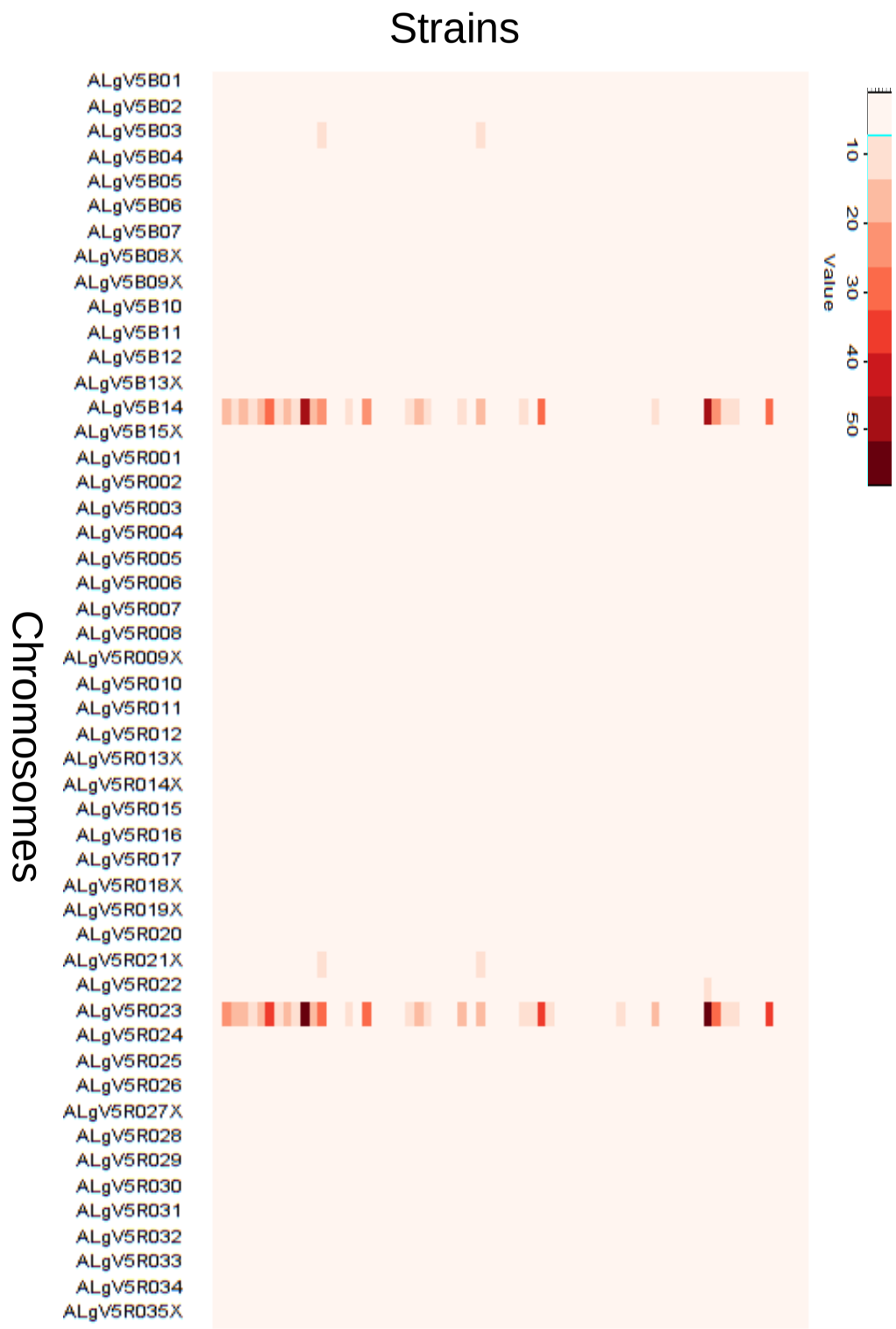


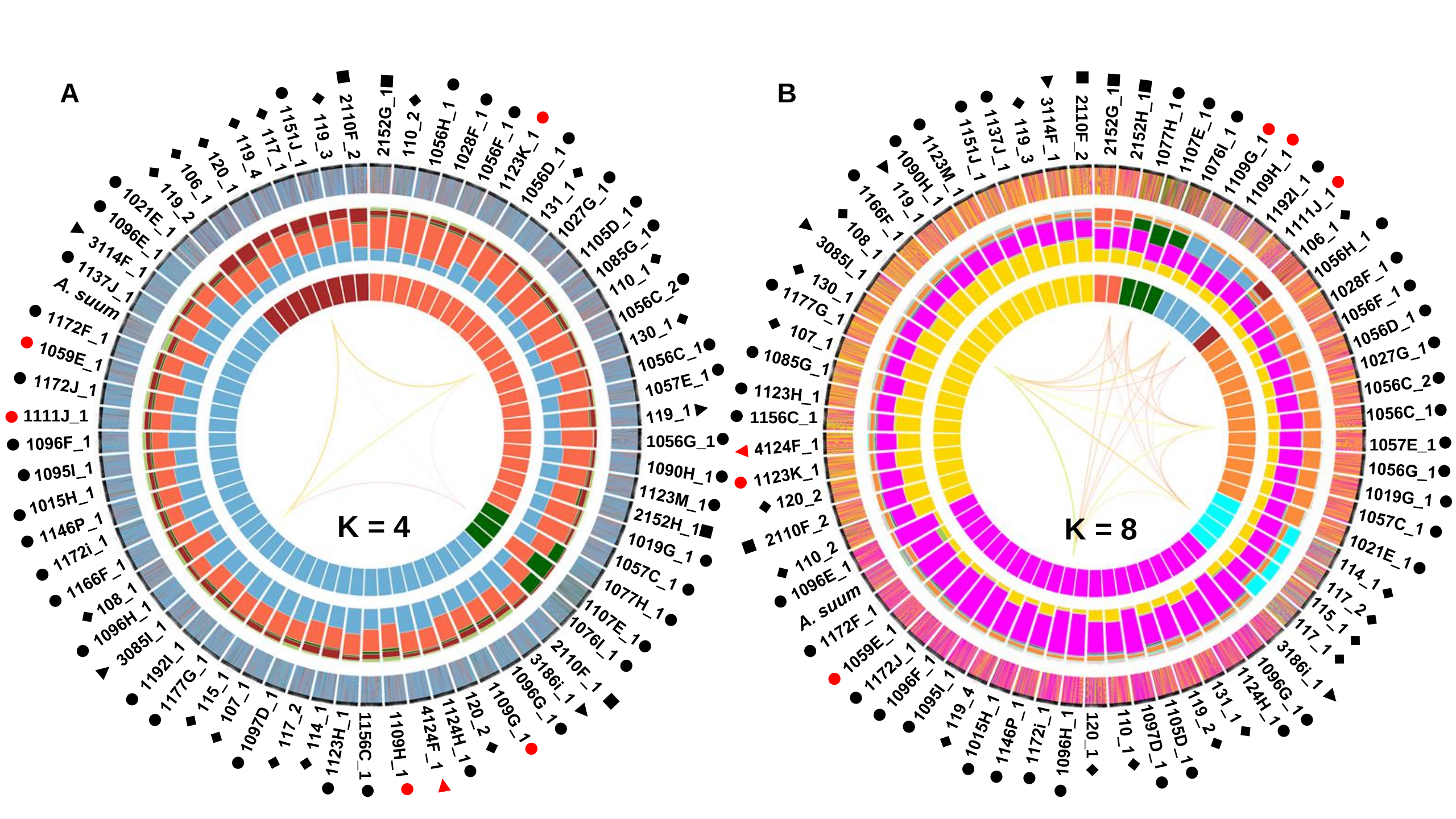
bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



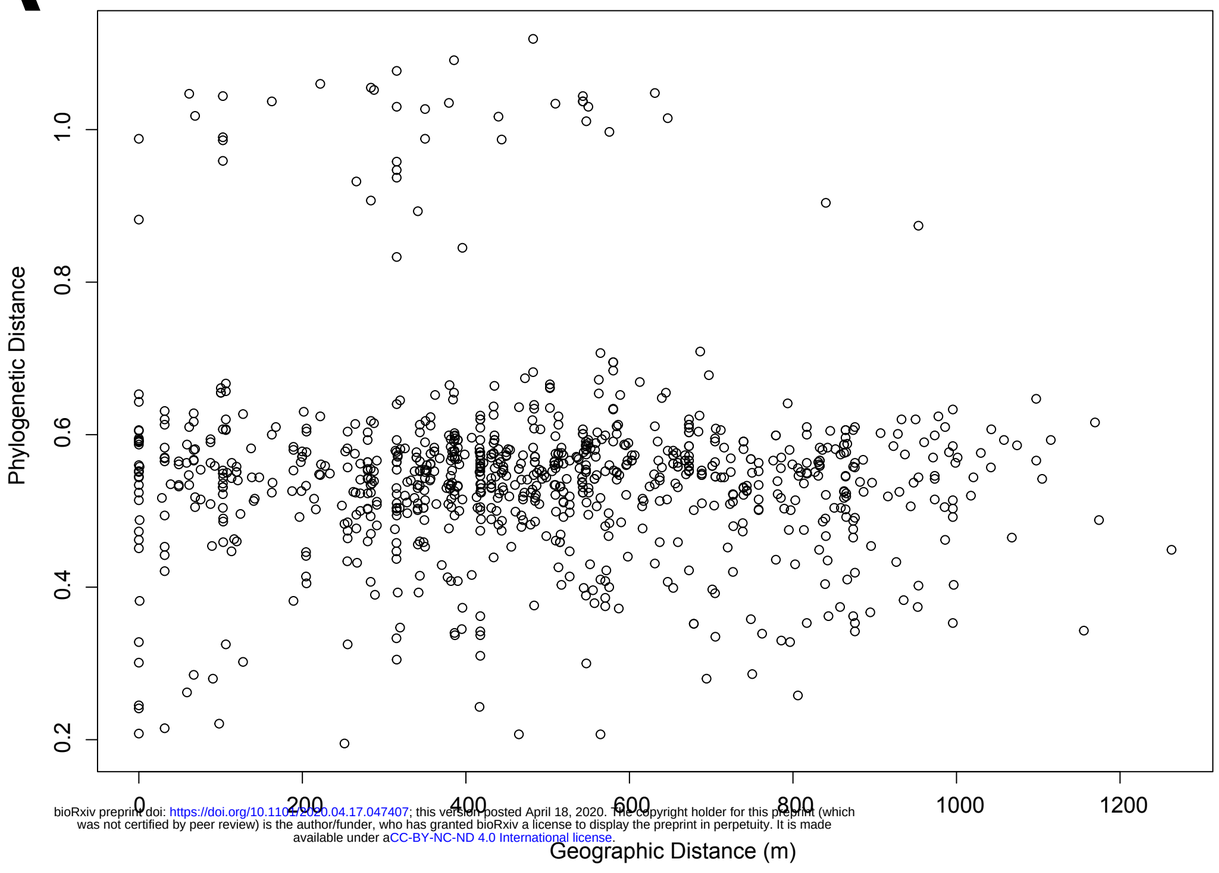


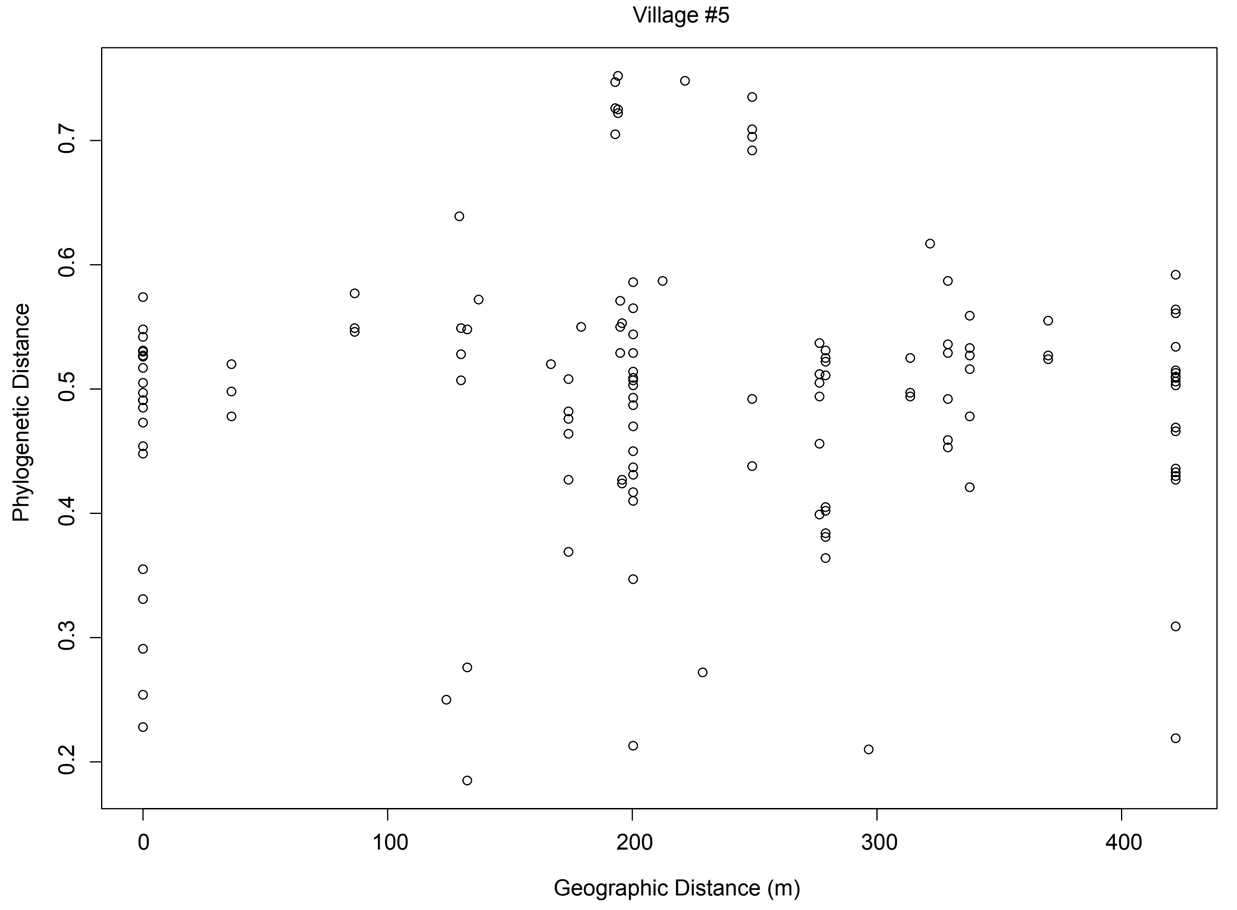



Strains
Chromosomes
ALgV5B01
ALgV5B02
ALgV5B03
ALgV5B04
ALgV5B05
ALgV5B06
ALgV5B07
ALgV5B08X
ALgV5B09X
ALgV5B10
ALgV5B11
ALgV5B12
ALgV5B13X
ALgV5B14
ALgV5B15X
ALgV5R001
ALgV5R002
ALgV5R003
ALgV5R004
ALgV5R005
ALgV5R006
ALgV5R007
ALgV5R008
ALgV5R009X
ALgV5R010
ALgV5R011
ALgV5R012
ALgV5R013X
ALgV5R014X
ALgV5R015
ALgV5R016
ALgV5R017
ALgV5R018X
ALgV5R019X
ALgV5R020
ALgV5R021X
ALgV5R022
ALgV5R023
ALgV5R024
ALgV5R025
ALgV5R026
ALgV5R027X
ALgV5R028
ALgV5R029
ALgV5R030
ALgV5R031
ALgV5R032
ALgV5R033
ALgV5R034
ALgV5R035X
Value
10
20
30
40
50


A
K = 4
2110F_2
2152G_1
110_2
1056H_1
1028F_1
1056F_1
1123K_1
1056D_1
131_1
1027G_1
1105D_1
1085G_1
110_1
1056C_2
130_1
1056C_1
1057E_1
119_1
1056G_1
1090H_1
1123M_1
2152H_1
1019G_1
1057C_1
1077H_1
1107E_1
1076I_1
2110F_1
3186i_1
1096G_1
1109G_1
120_2
1124H_1
4124F_1
1109H_1
1156C_1
1123H_1
114_1
117_2
1097D_1
107_1
115_1
1177G_1
1192I_1
3085I_1
1096H_1
108_1
1166F_1
1172i_1
1146P_1
1015H_1
1095I_1
1096F_1
1111J_1
1172J_1
1059E_1
1172F_1
A. suum
1137J_1
3114F_1
1096E_1
1021E_1
119_2
106_1
120_1
119_4
117_1
1151J_1
119_3
B
K = 8
2110F_2
2152G_1
2152H_1
1077H_1
1107E_1
1076I_1
1109G_1
1109H_1
1192I_1
1111J_1
106_1
1056H_1
1028F_1
1056F_1
1056D_1
1027G_1
1056C_2
1056C_1
1057E_1
1056G_1
1019G_1
1057C_1
1021E_1
114_1
117_2
115_1
117_1
3186i_1
1096G_1
1124H_1
131_1
119_2
1105D_1
1097D_1
110_1
120_1
1096I_1
1110H_1
1015H_1
1146P_1
1172i_1
119_4
1095I_1
1096F_1
1172J_1
1059E_1
1172F_1
A. suum
1096E_1
110_2
2110F_2
120_2
4124F_1
1123K_1
1156C_1
1123H_1
1085G_1
107_1
1177G_1
130_1
3085I_1
108_1
1166F_1
1090H_1
119_1
1123M_1
1151J_1
1137J_1
119_3
3114F_1

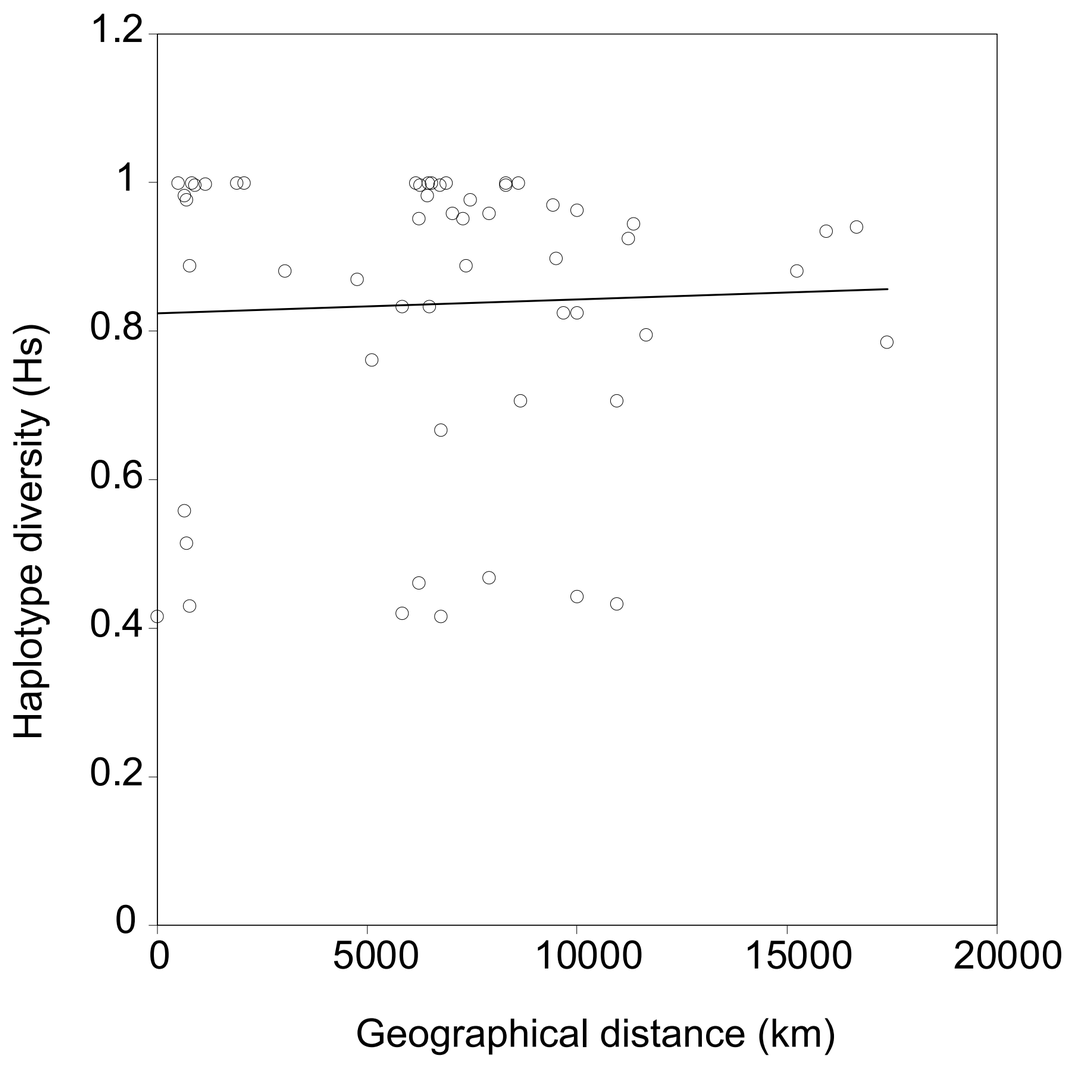
A

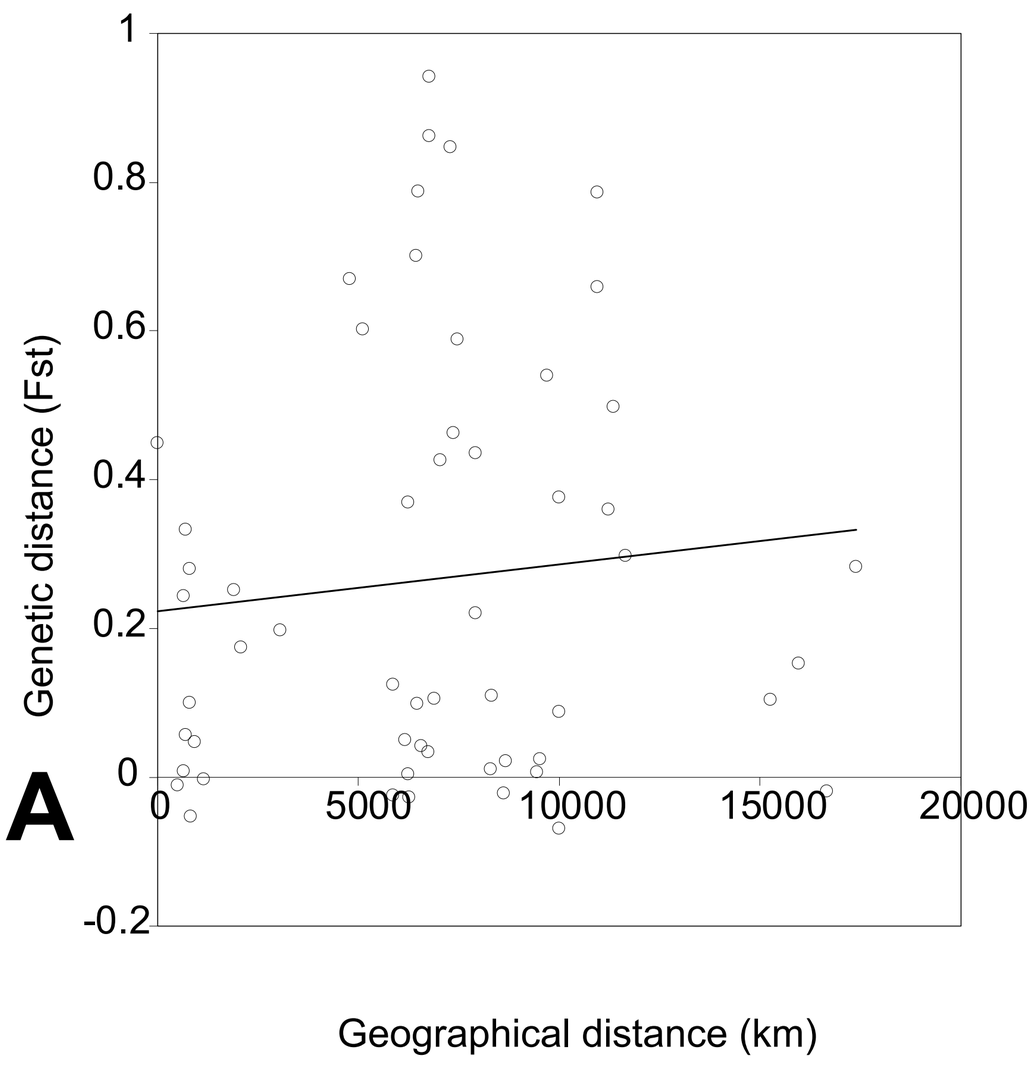




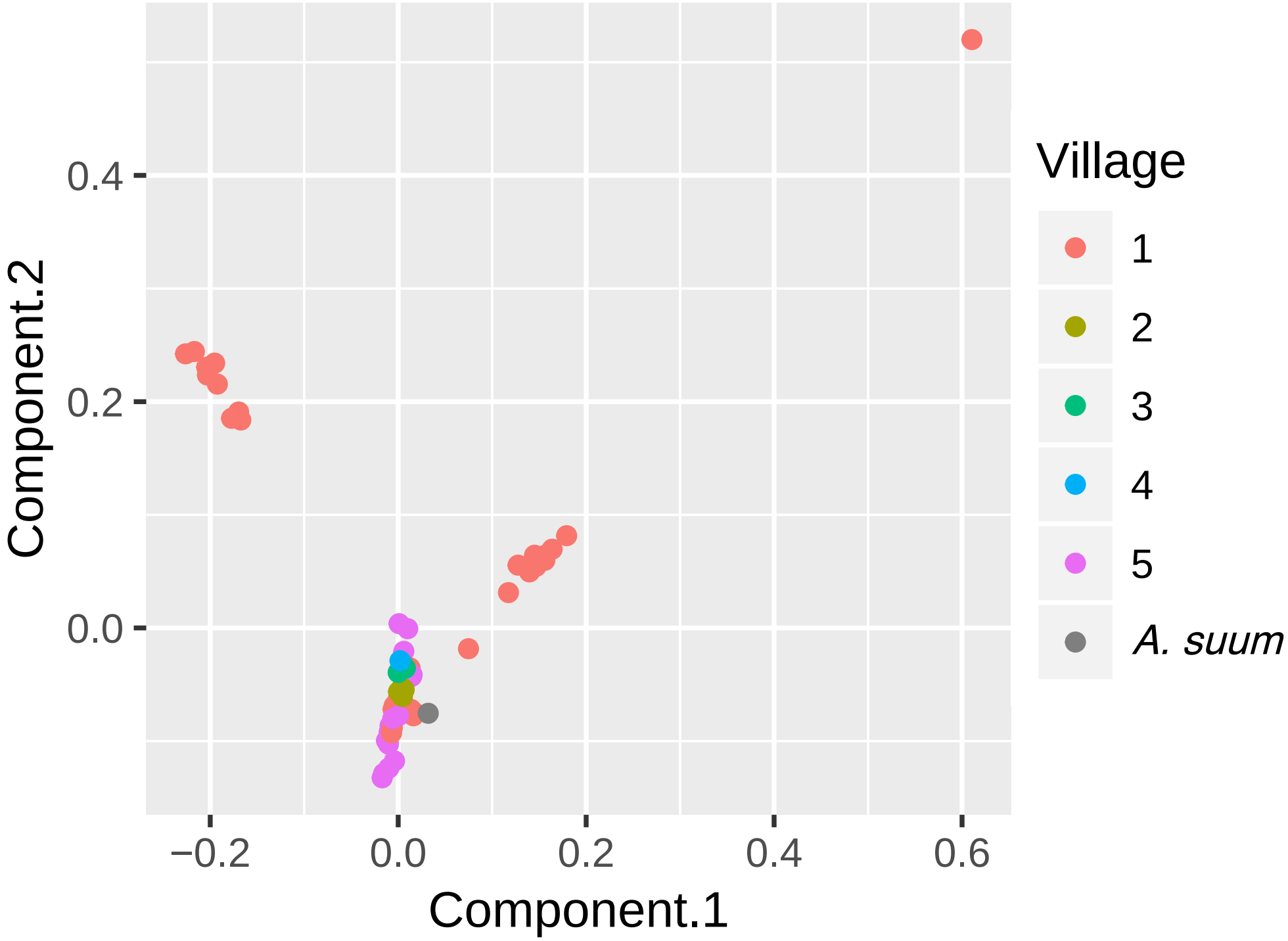
bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.047407; this version posted April 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

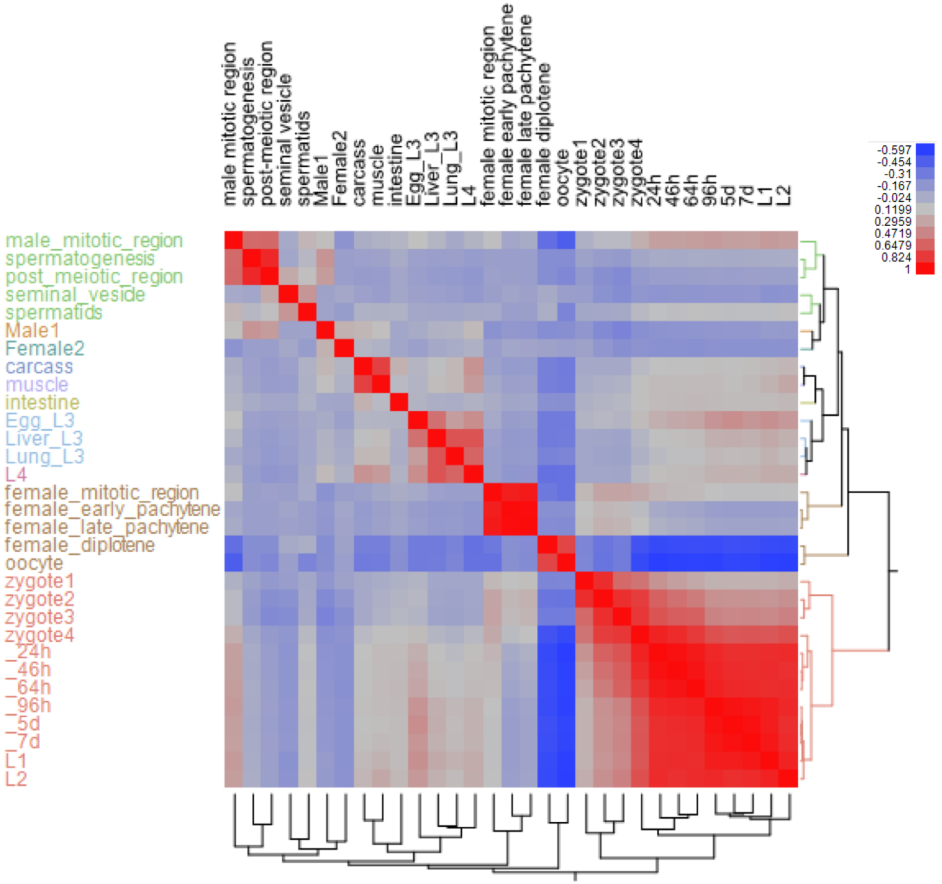
B

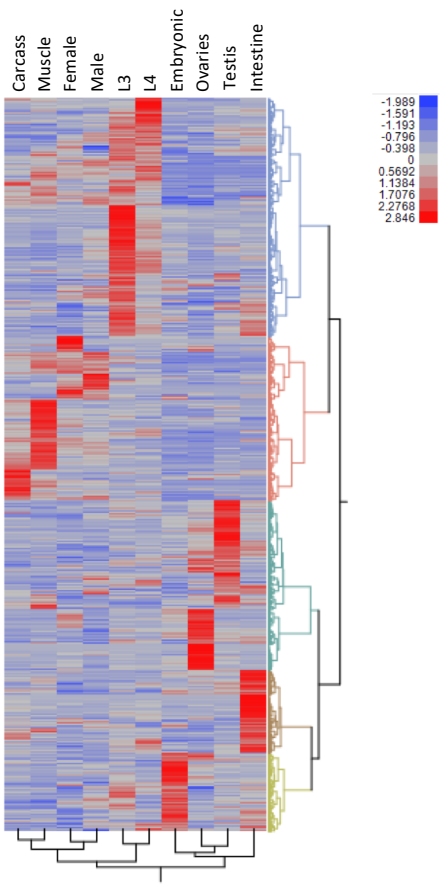


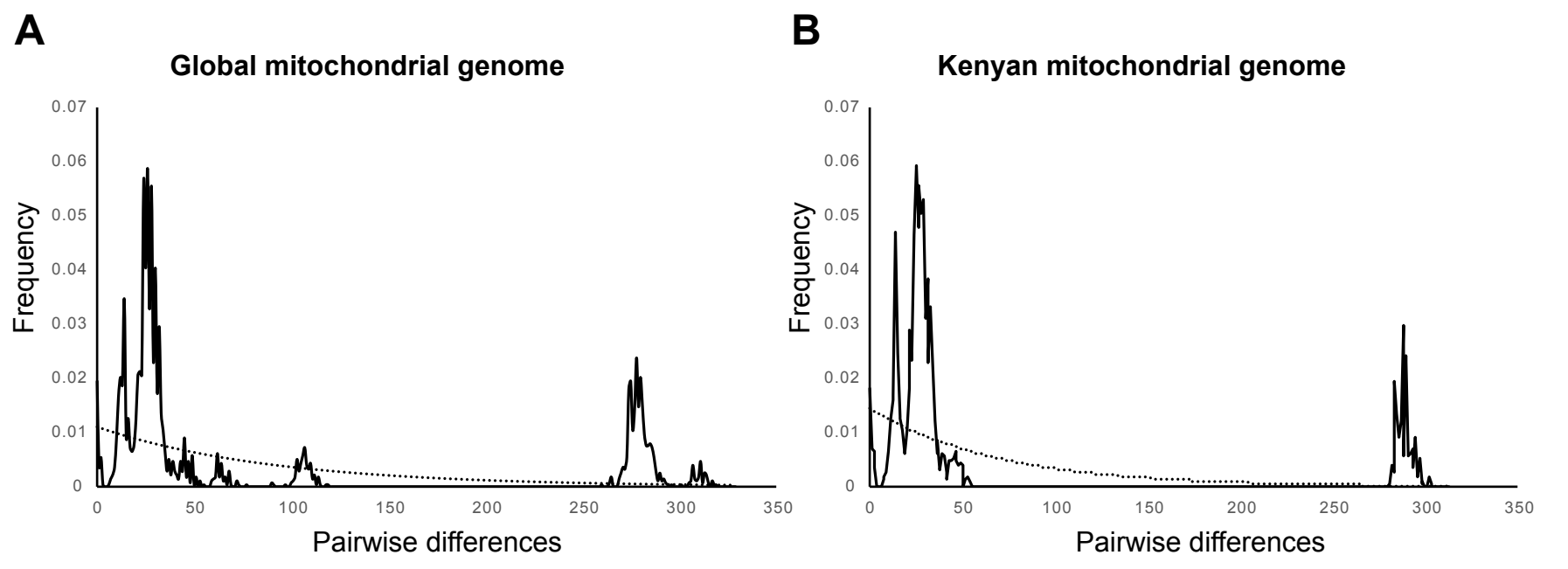


A


Component.2
0.4
0.2
0.0
−0.2
0.0
0.2
0.4
0.6
Component.1
Village
1
2
3
4
5
A. suum


male mitotic region
spermatogenesis
post-meiotic region
seminal vesicle
spermatids
Male1
Female2
carcass
muscle
intestine
Egg_L3
Liver_L3
Lung_L3
L4
female mitotic region
female early pachytene
female late pachytene
female diplotene
oocyte
zygote1
zygote2
zygote3
zygote4
24h
46h
64h
96h
5d
7d
L1
L2
-0.597
-0.454
-0.31
-0.167
-0.024
0.1199
0.2959
0.4719
0.6479
0.824
1


Carcass
Muscle
Female
Male
L3
L4
Embryonic
Ovaries
Testis
Intestine
-1.989
-1.591
-1.193
-0.796
-0.398
0
0.5692
1.1384
1.7076
2.2768
2.846


A
Global mitochondrial genome
Frequency
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
50
100
150
200
250
300
350
Pairwise differences
B
Kenyan mitochondrial genome
Frequency
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
50
100
150
200
250
300
350
Pairwise differences