Supplementary File S2. Technical notes for manuscript:

Machine learning-based analyses support the existence of species complexes for *Strongyloides fuelleborni* and *Strongyloides stercoralis*

Joel L. N. Barratt[1,2]*, Sarah G. H. Sapp[1]*

*corresponding authors

Email (S. G. H. Sapp): xyz6@cdc.gov

Email (J. L. N. Barratt): jbarratt@cdc.gov; joelbarratt43@gmail.com

[1] Centers for Disease Control and Prevention, Division of Parasitic Diseases and Malaria, Parasitic Diseases Branch

[2] Oak Ridge Associated Universities, Oak Ridge, Tennessee

Table of Contents

Supplementary Methods

Additional notes on the rationale for study design and the limitations of phylogeny

The ensemble of similarity-based classification algorithms employed here constitutes a machine learning (ML) procedure designed to address some limitations of phylogeny when assessing population structure by MLST (Table S1 - below). Differences in the way indels are treated by various multiple sequence aligners (e.g. ClustalW, MUSCLE) can impact downstream phylogenetic analyses with varying effects. For this reason, algorithms such as Gblocks (http://molevol.cmima.csic.es/castresana/Gblocks_server.html) are often used to remove regions of a sequence alignment containing gaps. This practice poses a problem for analysis of *Strongyloides* 18S rDNA sequences which possess informative indels that effectively differentiate genotypes (Figure 1, main text). Consequently, deletion of poorly aligned regions resulting from indels can result in misleading phylogenetic trees for these loci (Figure S1 - below).

Concatenation of multiple phylogenetically informative sequences from individual specimens is common practice when investigating phylogenetic relationships. However, aside from the fact that concatenating 18S rDNA and cox1 sequences of *Strongyloides* would not address the problem of indels (discussed above), this procedure also does not account for heterozygosity in diploid individuals. This becomes an increasing problem when multiple loci from a MLST study are heterozygous. In this study, 25 individual *S. stercoralis* were heterozygous at their 18S loci; in what way would you concatenate these sequences to your single cox1 haplotype? Finally, phylogenetic analysis cannot be performed on specimens that do not have data available for all targeted loci; partially typed specimens are typically excluded from a phylogenetic analysis. As published *Strongyloides* genotyping studies sometimes analyzed different marker combinations, combining these datasets for an all-inclusive phylogenetic analysis is not possible. The ML approach utilized here addresses these challenges (Table S1).

Table S1. Comparison of the ML approach to traditional methods for exploring taxonomic relationships

Common challenges/problems[α]	The ML approach used here[β]	Traditional approaches[γ]
MLST datasets may be incomplete due to failed sequencing at some loci or limitations on the amount of starting material available.	Missing data are tolerated extremely well. In this study, 619 of 837 specimens retained for analysis had data available for only the cox1 locus.	Sequences from multiple loci must be concatenated into a single sequence before analysis. If one or more of the MLST markers is missing for a specimen, that specimen is excluded from downstream analysis because a full-length concatenated sequence cannot be produced.
Some specimens are heterozygous at some nuclear loci included in a MLST panel.	This is not a confounding factor. The present study included 25 *S. stercoralis* specimens that were heterozygous at their 18S HVR-I and HVR-IV loci.	Traditional approaches do not address this problem. If a specimen is genotyped at 3 loci and each locus is heterozygous (e.g. possessing two alleles each), concatenation of sequences from this specimen would result in nine possible combinations. Heterozygous sites may be excluded to simplify the phylogenetic analysis, but this would result in a loss of critical information.
Loci possessing gaps, indels and repeat-based polymorphisms can be extremely informative but may confound traditional analysis approaches.	These genetic features are not a problem for the ML procedure. Haplotypes are defined by the user and whether they are based on SNPs, repeats, or indels, is irrelevant. *Strongyloides* 18S rDNA types possess each of these types of polymorphisms.	Phylogenetic methods and sequence clustering approaches typically ignore alignment gaps. Therefore, polymorphisms based on indels and repeats are generally ignored. When clustering by sequence similarity, gaps in the alignment can be considered in the similarity measure, though this is only helpful for short gaps. Considering large gaps in the similarity measure can lead to nonsensical results, as can poorly aligned regions containing indels. Phylogenetic trees generated from alignments of HVR-I and HVR-IV are shown in Figure S1, highlighting the impact indels can have on the results of a phylogenetic analysis.
Comparison of MLST data across multiple studies is difficult or impossible if the studies of interest use a slightly different panel of MLST markers.	A valid ML analysis can be performed across numerous studies if there is some overlap in the loci sequenced. Only 12% of specimens retained for analysis in this study had data available for all three markers.	In this study, cox1 sequence data was an absolute requirement for inclusion in our final analysis, so every specimen analysed in this study had cox1 data available. Therefore, a simple clustering analysis of cox1 sequences would be the most appropriate "traditional" analysis procedure that we could apply to this set of specimens for an all-inclusive comparison across these studies. A true phylogenetic analysis would not be advised on such short sequences (~217 bases); phylogeny attempts to reconstruct the evolutionary history of taxa and this is not advisable for a 217 base pair sequence. However, using our ML methodology, we integrated data from several MLST-based surveys that used different, but overlapping combinations of markers. Most specimens (74%) only had data available for cox1. However, the inclusion of even some specimens with 18S data contributes significantly to the final population structure when using the ML approach and improves the resolution of clusters (see Figures S2 and S3 below, pages 10 and 11).

[α] Refers to common problems and challenges that may be encountered when analyzing MLST datasets

[β] The machine learning approach used here refers to the normalized output of Barratt's heuristic classifier and Plucinski's naïve Bayes classifier, as described here: https://github.com/Joel-Barratt/Eukaryotyping

[γ] Includes phylogeny and other traditional methods such as clustering based on a genetic distance

Note: MLST means multi-locus sequence typing – refers to any procedure that involves sequencing multiple loci from individuals for the purposes of genotyping

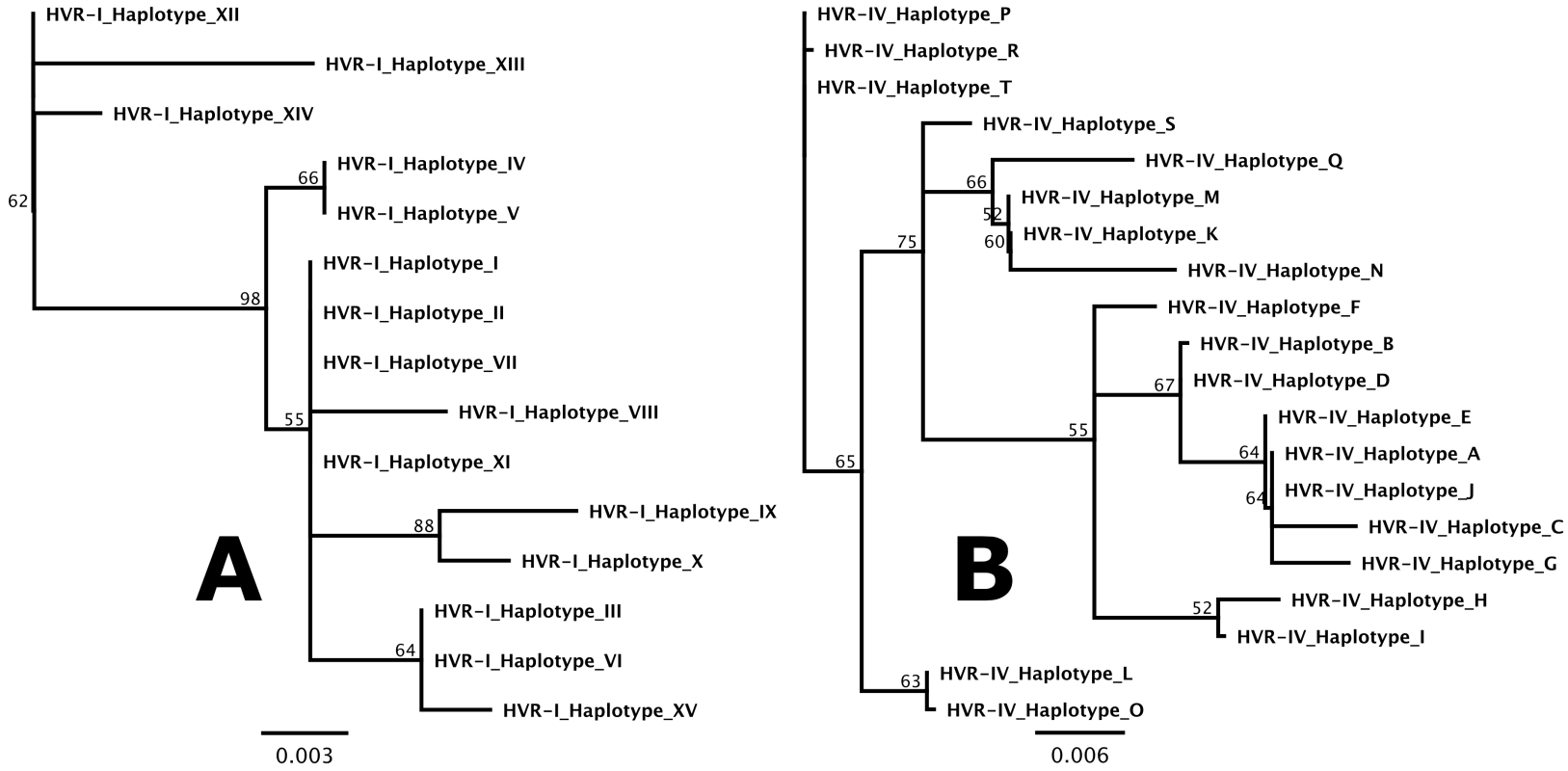


Figure S1. Phylogenetic analysis performed on 18S rDNA hypervariable region I (A) and hypervariable region IV (B) sequences of *Strongyloides*

Both trees were constructed using the Neighbor-Joining method and the Jukes-Cantor genetic distance model. Bootstrap values greater than 50% are shown on nodes (1000 bootstrap replicates). The tree in panel A considers 428 positions and that in panel B considers 196 positions. These trees are poor examples of phylogenetic analyses for several reasons, aside from the fact that the loci examined are probably too short for a robust phylogenetic analysis. In any case, every sequence included in these trees is distinct, but the sequences mainly differ by polymorphisms based on indels. As a consequence of these indels, the sequences align poorly at some locations (Figure 1, main manuscript text), leading to a loss of resolution during phylogenetic analysis. This phenomenon is exemplified by these trees which do not resolve some sequences well and possess poor bootstrap values at most nodes. In panel A for example, HVR-I haplotypes I, II and VII seem identical, yet an alignment of these sequences clearly demonstrates that they are not. Similarly, in panel B, HVR-IV haplotypes A and J seem identical but alignment of these sequences clearly shows that they differ by multiple indels. These trees were generated using Geneious Prime - Build 2018-11-0.

Generation of cox1 cluster dendrograms for comparison to ML cluster dendrograms

Construction of cox1 cluster dendrograms was performed to facilitate their direct comparison to dendrograms generated using ML for assessing the population structures of *S. fuelleborni* and *S. stercoralis*. Fasta sequence files containing all cox1 sequences (133 sequences for *S. fuelleborni* & *Strongyloides* sp. 'Loris', and 704 for *S. stercoralis*) were manually trimmed to the same length using Geneious software and exported as a fasta file. The resulting sequences were aligned using the 'msa' package in R (https://www.r-project.org/). A pairwise identity matrix was constructed using the 'seqinr' package 'dist.alignment' function. Clustering was then performed using agglomerative nested clustering 'agnes' with Manhattan distances and using the Ward clustering method. The 'ggtree' package was used to generate cluster dendrograms.

Notes on the preparation of sequence data for ML analysis

HVR-I and HVR-IV were divided into four and three fragments respectively, of approximately 100 bases (range of 75 to 111 bases), resulting in ~10 to 20 haplotypes for each of *S. fuelleborni* & *Strongyloides* sp. "Loris" and *S. stercoralis* at the most variable segments of these loci (Table S2 - below). Cox1 sequences were divided into nine segments so that each segment possessed approximately 20 haplotypes; ~10 haplotypes each for the *S. stercoralis* and the *S. fuelleborni* & *Strongyloides* sp. "Loris" datasets. These nine cox1 segments are defined as A1 to A3, B1 to B3, and C1 to C3 (Table S2 - below). These divisions were empirically assigned to ensure that the number of haplotypes at any given segment was approximately 10. Based on extensive testing of the ML-based approach, loci possessing more than ~20 haplotypes may limit the algorithms ability to identify genetic links at these loci, and this effect increases steadily with the number of haplotypes that exist at a given locus. This phenomenon is observed because loci with an excessively high haplotype diversity are assigned a heavier weight by the algorithm when

calculating the relative distance between specimens in the study population. As a consequence of this, if the haplotype diversity is too high, the number of shared haplotypes between members of the population is low, yet the locus will still contribute greatly to the algorithm's prediction of the population structure due to its high diversity. The net result is that the population might appear less structured (or unstructured) when the output of the ML algorithm is visualized. Note that this is a generalization based on some populations that we have tested (other examples to be published elsewhere), and this effect does vary. In any case, splitting loci with an excessively high haplotype diversity into smaller segments to limit the number of haplotypes at each segment is a viable solution for extracting valuable information from loci with an excessively high-entropy (Table S2 - below). In this study the authors made these divisions empirically, but the developers of the ML method (https://github.com/Joel-Barratt/Eukaryotyping) are writing software that performs this task automatically in an optimized fashion for any dataset.

Table S2: Division of genotyping markers into smaller segments for ML analysis

Original locus	Segments	Length of segments (bases)	Number of haplotypes per segment*
HVR-I	Part A	99 - 104	12
	Part B	111	3
	Part C	110	4
	Part D	100	3
HVR-IV	Part A	87	3
	Part B	75 - 87	17
	Part C	85	4
COX1	Part A1	25	19
	Part A2	25	18
	Part A3	20	13
	Part B1	25	22
	Part B2	25	22
	Part B3	20	18
	Part C1	25	21
	Part C2	25	16
	Part C3	27	16

Note: The sequence of all haplotypes is provided in **Appendix A** of this document

*Includes haplotypes of both the *S. stercoralis* and *S. fuelleborni* & *Strongyloides* sp. "Loris" datasets.

Dividing markers into segments was a helpful practice for the very short (217 base pair) cox1 region examined here which possesses more than 100 distinct haplotypes leading to extreme resolution and a loss of structure. However, there are some caveats to consider prior to dividing markers for the purposes of obtaining ~10 haplotypes at each segment. Firstly, if meeting this requirement leads to segments that are excessively short (~10 bases or less), splitting haplotypes is not recommended because this could result in a loss of specificity because shorter sequences (e.g. k-mers) are more likely to appear in multiple places in a sequence dataset. In this study, we used a BLASTN similarity search to identify the haplotypes in each specimen. If the subject sequences used in a BLASTN search are 10 bases or less, this greatly increases the risk of these sequences matching to unrelated sequences (i.e. false positive haplotype detection). In the case of cox1, our haplotypes were kept between 20 and 27 bases long which avoids this problem for the present dataset and conveniently kept the number of haplotypes at ~10 at each segment, for both the *S. stercoralis* and *S. fuelleborni* & *Strongyloides* sp. "Loris" datasets (Table S2 – above).

This issue of dividing loci also raises another point for consideration; whether markers with excessive haplotype diversity are worth considering in population level studies. Loci with excessively high diversity are more likely to be evolutionarily neutral and rapidly mutating, resulting in high rates of sequence divergence. Sequences possessing these characteristics possess limited value for identifying population-level trends and population geneticists should be cognizant of this prior to selecting a genotyping panel.

Additionally, obtaining approximately 10 haplotypes for all loci is not a hard requirement for the ML procedure and there is certainly some flexibility in this value. For example, a haplotype numbers less than 10 and greater than 30 will still yield meaningful results in most circumstances. For example, if a study population includes only 60 individuals, and one marker in your MLST

panel possesses 60 haplotypes (one unique type per individual), then this marker is useless for identifying population trends among members of this group. One could split this marker into segments (as performed in this study for cox1) in an attempt to address this, or exclude the marker altogether if dividing markers is not feasible due to the caveats discussed above. Another example; if in this same population of 60, a second marker in the MLST panel possesses 30 haplotypes, it will probably still provide some value, particularly if there are other markers in the selected MLST panel with less diversity (e.g. closer to 10) to compensate. Alternatively, if a marker possesses 30 haplotypes among a population of 1,000 individuals, this is probably an excellent marker for this population. Flexibility in the number of haplotypes required for each locus is important to note, because if novel haplotypes are discovered at a later stage during an MLST study, users of this method are certainly not required to divide markers into new segments (or identify new markers) in order to retain a haplotype number of ~10 per segment.

Another benefit of dividing markers into segments is that this can remove the problem of PCR-induced chimera formation. PCR-induced chimeras are artefacts produced when two true haplotypes generate hybrids of each other *in vitro*. This occurs when a partial PCR product from one haplotype incorrectly primes the extension of a different haplotype. This is a common artefact observed in deep-amplicon sequencing experiments that can be a major problem for studies aiming to identify novel haplotypes. These artefacts appear (and are continually amplified) prior to deep sequencing, and therefore seem “real”. These artefacts cannot be removed by quality trimming or quality filtering. Dividing sequences into segments where each segment represents a distinct variable site, can greatly mitigate (or completely eliminate) the impact of these artefacts.

Selection of the minimum data requirements for classification

The ML approach does not require a complete set of markers for a robust analysis. While this is an important benefit over traditional approaches such as phylogeny, some minimum data requirements must still be set by the user. This is because a researcher cannot expect this procedure to accurately classify a specimen in the event that sequence data is generated for only 20 bases of a single MLST marker for example. When the user is establishing these requirements, the most informative markers (i.e. based on their entropy or based on knowledge available in the scientific literature) – are generally a good start. These data requirements are an adjustable user-defined parameter and it is up to the user to experiment with these in order to establish optimal settings that address their specific question (https://github.com/Joel-Barratt/Eukaryotyping). This can be done empirically, based on the scientific literature, or by assessing algorithm performance experimentally on various test datasets (obviously, the latter procedure is preferred).

In the present study we set two minimum data requirements. Firstly, *Strongyloides* cox1 sequences are hypervariable, possessing high haplotype diversity. Consequently, this locus is very informative from a genotyping standpoint. Similarly, the 18S HVR-IV locus alone can distinguish the two major lineages of *S. stercoralis* (A and B) based on evidence from the scientific literature. These observations led us to establish that firstly, published *Strongyloides* sp. genotypes had to include all 9 cox1 segments (Table S2 - above) if cox1 was the only sequence available. Secondly, if all 9 cox1 fragments were not available, the ML procedure was set to tolerate as few as 7 out of 9 of these segments, but in this case, HVR-IV Part B (Figure 1, main manuscript text) was also required to compensate for the truncated cox1 sequence. Specimens that did not meet these criteria were excluded from the analysis. If HVR-I was available for a given specimen its sequence was also included to guide the ML procedures classifications, but the availability of HVR-I was not an

absolute requirement. Instructions on how to adjust these settings and the R-scripts required to run this ML analysis are provided here: https://github.com/Joel-Barratt/Eukaryotyping.

Supplementary Results and Discussion

Generation of cox1 cluster dendrograms for comparison to our ML approach

The ML cluster dendrogram (Figure S2, panel A) and cox1 cluster dendrogram (Figure S2, panel B) constructed for *S. stercoralis* reflect a similar population structure. However, the ML approach provides additional resolution as we observe an additional cluster (Figure S2, panel A). Furthermore, six specimens known to belong to the B lineage of *S. stercoralis* based on their 18S HVR-IV data were placed in a cluster alongside specimens of *S. stercoralis* lineage A using the standard cox1 clustering method (Figure S2, panel B, cluster 2). The relationships represented in the *S. fuelleborni* cox1 dendrogram (Figure S3, panel B) are also very similar to those depicted in the ML dendrogram (Figure S3, panel A). However, by clustering only cox1 sequences (Figure S3, panel B), *S. fuelleborni* from Tanzania (Figure S3, panel B, cluster 1) were positioned closely to *S. fuelleborni* from Japanese macaques (Figure S3, panel B, cluster 4). However, the ML procedure (Figure S3, panel A) places cluster 3 (*S. fuelleborni* from Southeast Asian macaques) nearest to cluster 4 (*S. fuelleborni* from Japanese macaques), which is a more plausible classification. Clusters 3 and 4 each include *S. fuelleborni* collected exclusively from Asian *Macaca* species, and this observation is consistent with what is logically expected, and suggests allopatric speciation. We also know that *S. fuelleborni* from clusters 3 and 4 each possess HVR-IV haplotype S, suggesting they do indeed share a closer relationship to each other than to *S. fuelleborni* from Africa, which possess HVR-IV haplotypes P, Q, K or L.

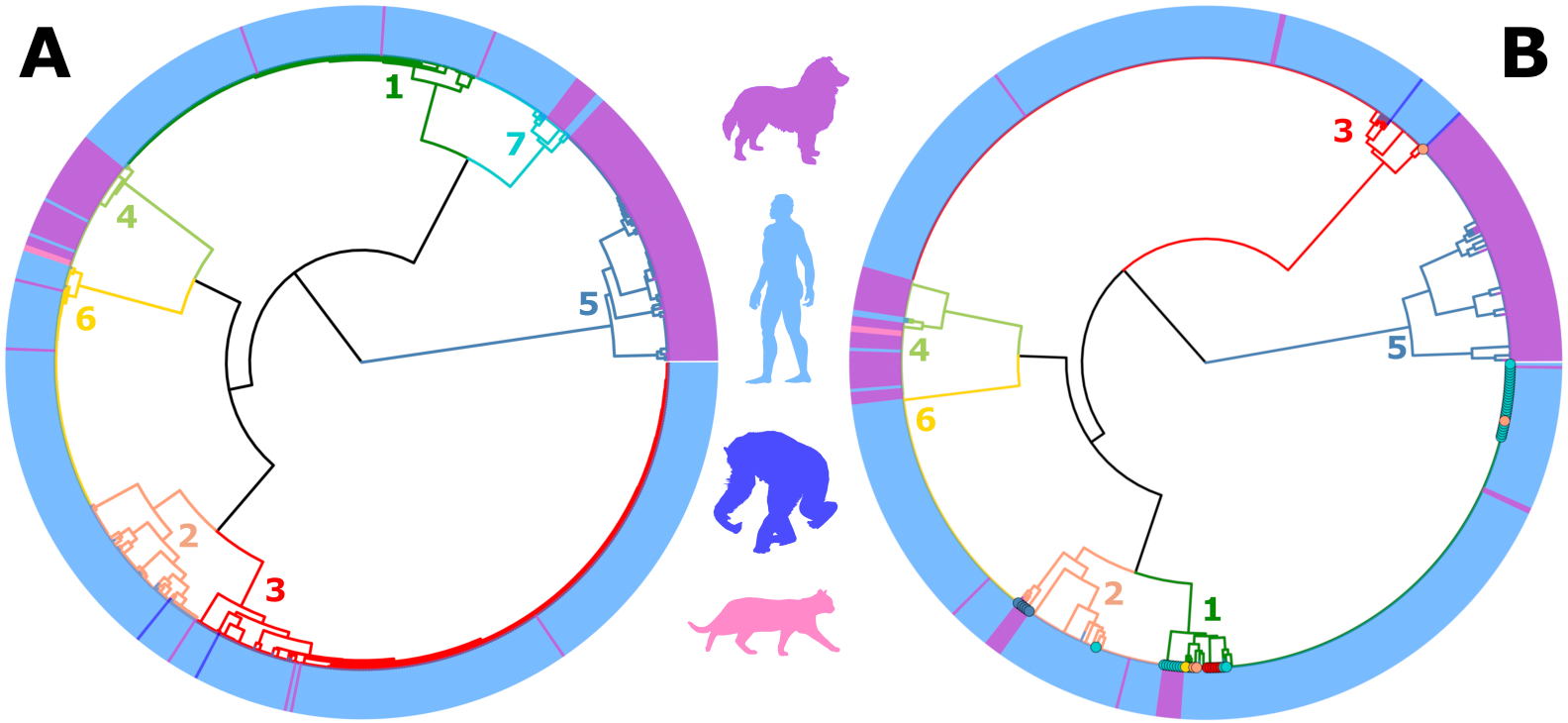


Figure S2. Population structure of *S. stercoralis* predicted using ML (A) versus that predicted by clustering of cox1 sequences only (B)

The ML dendrogram in panel A is identical to Figure 2 (main manuscript text), while that shown in panel B was generated from the same *S. stercoralis* specimens included in panel A, though by clustering of their cox1 sequences based on sequence similarity only. Branches in panel B are color coded according to the clusters shown in panel A, as determined by ML. Peripheral bars are color coded according to the host from which the specimen was derived. To include all cox1 sequences in panel B that were analyzed in panel A, the cox1sequences had to be trimmed to 170 bases because traditional sequence clustering requires that all sequences are of the same length (panel B). Colored dots were placed on branch tips (panel B) for specimens that were not assigned to the same cluster by ML (Panel A). The dots are color coded according to the cluster numbers in Panel A, based on the clusters they were assigned to by ML. Clustering of cox1 sequences alone resulted in a loss of resolution (cluster 7 is not visible in panel B), and the assignment of some specimens to incorrect clusters. For example, six specimens assigned to cluster 5 using ML were assigned to cluster 2 (panel B, navy blue dots). Each of these specimens (navy dots) possesses HVR-IV haplotype B and specimens with this haplotype were exclusively assigned to cluster 5 using ML.

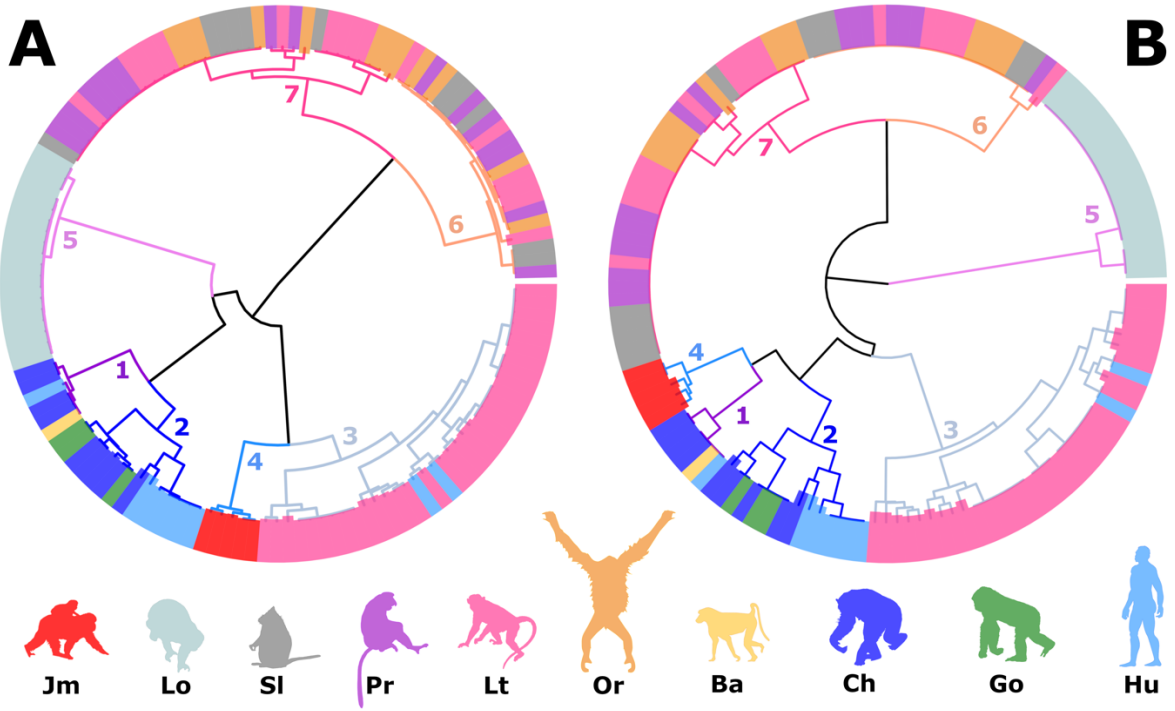


Figure S3. Population structure of *S. fuelleborni* predicted using machine learning (A) versus that predicted by clustering of cox1 sequences only (B)

The ML dendrogram (panel A) is identical to Figure 3 (manuscript text). The dendrogram shown in panel B was generated from the same *Strongyloides* specimens as panel A, though by clustering cox1 sequences by similarity only. Branches in panel B are color coded according to the clusters in panel A, as determined by ML. Peripheral bars are color coded according to the host from which the specimen was derived; either human (Hu), gorilla (Go), chimpanzee (Ch), baboon (Ba), orangutan (Or), long-tailed macaque (Lt), proboscis monkey (Pr), silvered leaf monkey (Sl), slow loris (Lo), or Japanese macaque (Jm). To include the same specimens in both analyses (A and B), the cox1 sequences in panel B had to be trimmed to 202 bases because sequence clustering requires all sequences to be the same length. The structure of these dendrograms is similar, with the exception that the position of cluster 4 (containing *S. fuelleborni* from Japanese macaques) is different. In panel B, cluster 4 is placed in a position nearest to *S. fuelleborni* from Tanzania (cluster 1). In panel A, cluster 4 is placed nearest to *S. fuelleborni* from SE Asian long-tailed macaques (cluster 3). The ML assignment is more plausible because clusters 3 and 4 each include *S. fuelleborni* collected exclusively from Asian *Macaca* species. Additionally, *S. fuelleborni* from Japanese macaques and long-tailed macaques each possess HVR-IV haplotype S which is distinct from HVR-IV haplotypes from all African *S. fuelleborni*.

Notes on the number of specimens with an incomplete genotype retained for analysis

If we wished to analyze as many specimens as possible in the present analysis using phylogenetic methods, and we required that both 18S and cox1 sequences were available, we would have to exclude all specimens lacking 18S data – for this dataset 88% of all specimens would be excluded. Similarly, if we wished to build a phylogeny only for *S. fuelleborni*, and we wished to exclusively analyze specimens with HVR-I, HVR-IV and cox1 sequences available, only 5 of 133 specimens could be examined. A major benefit of the ML approach is that specimens with a partial genotype can be retained for analysis. Table S3 shows the markers available for specimens that met the minimum data requirements for inclusion in this analysis. Note that the vast majority of these had only a partial genotype available, yet this additional information resulted in an increase in resolution over what traditional methods provide (Figures S2 and S3), while avoiding exclusion of a huge number of specimens from this analysis.

Table S3: Data completeness metrics for specimens included in the analysis

	Only Cox1	Only Cox1 & HVR-IV	Only Cox1 & HVR-I	All markers	Total
S. stercoralis	533 (75.7%)	64 (9.1%)	11 (1.6%)	96 (13.6)	704 (100%)
S. fuelleborni and *Strongyloides* sp. "Loris"	86 (64.7%)	41 (30.8%)	1 (0.8%)	5 (3.8%)	133 (100%)
Total	619 (74%)	105 (12.5%)	12 (1.4%)	101 (12.1%)	837 (100%)

Notes on potential sampling biases

The aim of this study was to analyze large extant MLST datasets available for *S. stercoralis* and *S. fuelleborni* by combining data from multiple studies in an attempt to elucidate relationships among genotypes, their hosts, and geographic origins beyond what has been shown when data from individual studies are analyzed individually. Any patterns that emerged as a result of this

analysis we would assess the significance of using a simple chi-squared test. While this proved helpful in supporting the distribution of specific *S. stercoralis* types in different hosts, statistical tests were not always applied, sometimes because the data alone were self-supporting (e.g. 98% of *S. stercoralis* assigned to cluster 5 came from dogs infected in SE Asia, or 19 of 20 *S. fuelleborni* assigned to clusters 1 and 2 came from Africa). In other cases, we did not apply statistical tests to explore possible trends due to the potential impacts of sampling bias. The study design involved an unbiased selection of all available *S. stercoralis* and *S. fuelleborni* sequences in GenBank at the time of analysis regardless of their host and geographic origin. Despite this, we did observe over-sampling from some hosts and certain geographic locations (Tables S4 to S7).

Table S4: Frequency matrix displaying country of origin versus host for *S. stercoralis*

	Humans	Dogs	Cats	Chimpanzees	TOTAL
Australia	10	1	0	0	11
Brazil	3	0	0	0	3
Cambodia	11	16	0	0	27
Central African Republic	3	0	0	0	3
China	6	0	0	0	6
Guinea	1	0	0	0	1
Iran	10	0	0	0	10
Italy	1	2	0	0	3
Ivory Coast	1	0	0	0	1
Japan	88	34	0	1	123
Laos	41	0	0	0	41
Myanmar	243	83	0	0	326
Nigeria	3	0	0	0	3
Switzerland	0	3	0	0	3
Tanzania	1	0	0	1	2
Thailand	124	0	0	0	124
Uganda	8	0	0	0	8
USA	0	7	0	0	7
West Indies	0	0	2	0	2
TOTAL	554	146	2	2	704

Note: Boxes containing numeric values are shaded according to their frequency, with the highest frequencies shown in black and the lowest frequencies in white.

The *S. stercoralis* sampled includes those from all inhabited continents, but the vast majority came from humans infected in Myanmar, and Southeast (SE) Asia more generally (Table S4 – above). *S. stercoralis* collected from humans in Japan are also well sampled. *S. stercoralis* infecting dogs from Japan and SE Asia are also far better sampled than dogs from other parts of the world (Table S4). Additionally, *S. stercoralis* causing human infections are better sampled than those causing infections in dogs regardless of their geographic origin, yet the study still samples are relatively large number of dogs (n = 146), which constitutes 20.7% of all *S. stercoralis* specimens analyzed here. Consequently, for *S. stercoralis* we focus on substantiating the relationship between genotype and host, yet we observe some patterns relating to geography that are worth noting when the dataset is dissected by genetic cluster (Table S5 – below).

Table S5: Frequency matrix displaying cluster assignment versus country of origin for *S. stercoralis*

	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5	Cluster_6	Cluster_7	TOTAL
Australia	1	3	1	0	1	0	5	11
Brazil	0	0	2	0	0	0	1	3
Cambodia	0	4	6	0	13	4	0	27
Central African Republic	0	0	0	0	0	0	3	3
China	0	6	0	0	0	0	0	6
Guinea	0	0	0	0	0	0	1	1
Iran	0	3	6	0	0	1	0	10
Italy	0	0	1	2	0	0	0	3
Ivory Coast	0	1	0	0	0	0	0	1
Japan	83	4	2	25	1	0	8	123
Laos	2	6	23	2	0	8	0	41
Myanmar	54	9	132	0	77	42	12	326
Nigeria	0	0	0	0	0	0	3	3
Switzerland	0	0	0	3	0	0	0	3
Tanzania	0	2	0	0	0	0	0	2
Thailand	0	20	59	2	0	35	8	124
Uganda	0	0	8	0	0	0	0	8
USA	0	0	0	7	0	0	0	7
West Indies	0	0	0	2	0	0	0	2
TOTAL	140	58	240	43	92	90	41	704

Note: Boxes containing numeric values are shaded according to their frequency, with the highest frequencies shown in black and the lowest frequency (zero) in white.

A total of 92 *S. stercoralis* were assigned to genetic cluster 5 and 90 of these (98%) came from SE Asia; either Cambodia (n = 13) or Myanmar (n = 77). Each of these 92 infections (100%) occurred in a dog. This suggests that a canine-specific variety of *S. stercoralis* (*S. stercoralis canis*?) originates from and is endemic to SE Asia (Table S5). *S. stercoralis* assigned to other genetic clusters seem to have a more cosmopolitan distribution, although hypotheses surrounding origin and endemicity of other *S. stercoralis* types cannot be proposed in the absence of improved sampling of dogs and humans from other locations.

Regarding *S. fuelleborni*, long-tailed macaques are well sampled, particularly those from mainland SE Asia, but long-tailed macaques from Malaysia are also comparatively well sampled (Table S6 – below). Other host species from other locations are poorly sampled by comparison, so discussion of the relationships between the various *S. fuelleborni* types focuses mostly on geography, for which the data are relatively self-supporting in the absence of additional statistical support. Firstly, *S. fuelleborni* specimens assigned to the closely related clusters 1 and 2 (Figure S3), predominantly include worms from Africa. In fact, 19 of the 20 *S. fuelleborni* specimens assigned to these two clusters (95%) originate from Africa (Table S7 - below). All specimens (100%) obtained from SE Asian long-tailed macaques (n = 34) were assigned to cluster 3 along with an infection that occurred in an Indian human. The *S. fuelleborni* types assigned to cluster 3 seem closely related to *S. fuelleborni* from Japan that were assigned to cluster 4 (n = 5), and distantly related to the African types (clusters 1 and 2). Certainly, the *S. fuelleborni* types from Malaysia (Clusters 6 and 7) are distinct to those from both Africa and mainland SE Asia, and show no clustering based on host preference. Similarly, the *Strongyloides* sp. from Malaysian slow lorises (cluster 5) is distinct (Table S7).

Table S6: Frequency matrix of *Strongyloides fuelleborni* and *Strongyloides* sp. country of origin versus host

	Human	Chimpanzee	Baboon	Gorilla	Long-tailed macaque	Japanese macaque	Bornean slow loris	Orangutan	Proboscis monkey	Silvered leaf monkey	TOTAL
Tanzania	1	4	1	0	0	0	0	0	0	0	6
Gabon	0	2	0	1	0	0	0	0	0	0	3
Central African Republic	5	3	0	2	0	0	0	0	0	0	10
Guinea-Bissau	1	0	0	0	0	0	0	0	0	0	1
India	1	0	0	0	0	0	0	0	0	0	1
Thailand	1	0	0	0	25	0	0	0	0	0	26
Laos	0	0	0	0	8	0	0	0	0	0	8
Japan	0	0	0	0	0	5	0	0	0	0	5
Malaysia	0	0	0	0	16	0	18	12	16	11	73
TOTAL	9	9	1	3	49	5	18	12	16	11	133

Note: *Strongyloides* sp. refers to the undescribed species previously detected from the Bornean slow loris. Boxes containing numeric values are shaded according to their frequency, with the highest frequencies shown in black and the lowest frequency (zero) in white.

Table S7: Frequency matrix of *S. fuelleborni* & *Strongyloides* sp. from various hosts from different countries versus cluster (A) and country of origin versus cluster (B)

	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5	Cluster_6	Cluster_7	TOTAL	
Humans from Tanzania	1	0	0	0	0	0	0	1	
Chimpanzees from Tanzania	4	0	0	0	0	0	0	4	
Baboons from Tanzania	1	0	0	0	0	0	0	1	
Chimpanzees from Gabon	0	2	0	0	0	0	0	2	
Gorillas from Gabon	0	1	0	0	0	0	0	1	
Chimpanzees from the Central African Republic	0	3	0	0	0	0	0	3	
Gorillas from the Central African Republic	0	2	0	0	0	0	0	2	
Humans from the Central African Republic	0	5	0	0	0	0	0	5	
Humans from Guinea-Bissau	0	1	0	0	0	0	0	1	
Humans from India	0	0	1	0	0	0	0	1	**A**
Humans from Thailand	0	0	1	0	0	0	0	1	
Long-tailed macaques from Thailand	0	0	25	0	0	0	0	25	
Long-tailed macaques from Laos	0	0	8	0	0	0	0	8	
Japanese macaques	0	0	0	5	0	0	0	5	
Bornean slow loris from Malaysia	0	0	0	0	18	0	0	18	
Orangutans from Malaysia	0	0	0	0	0	4	8	12	
Long-tailed macaques from Malaysia	0	0	0	0	0	6	10	16	
Proboscis monkeys from Malaysia	0	0	0	0	0	7	9	16	
Silvered leaf monkeys from Malaysia	0	0	0	0	0	5	6	11	
TOTAL	6	14	35	5	18	22	33	133	
Tanzania	6	0	0	0	0	0	0	6	
Gabon	0	3	0	0	0	0	0	3	
Central African Republic	0	10	0	0	0	0	0	10	
Guinea-Bissau	0	1	0	0	0	0	0	1	
India	0	0	1	0	0	0	0	1	**B**
Thailand	0	0	26	0	0	0	0	26	
Laos	0	0	8	0	0	0	0	8	
Japan	0	0	0	5	0	0	0	5	
Malaysia	0	0	0	0	18	22	33	73	

Note: *Strongyloides* sp. refers to the undescribed species previously detected from the Bornean slow loris. Boxes containing numeric values are shaded according to their frequency, with the highest frequencies shown in black and the lowest frequency (zero) in white.

Appendices

Appendix Part A. Fasta sequences of each *Strongyloides* haplotype included in this analysis

```
>COXI_PART_A1_Hap_1_
TTTGATTTATCAGCATTTGTTTTGA
>COXI_PART_A1_Hap_2_
TTTGGTTTATCAACATTTGTTTTGG
>COXI_PART_A1_Hap_3_
TTTAATTTATCAGCATTTATTCTGA
>COXI_PART_A1_Hap_4_
TTTAATTTATCAACATTTATTCTGA
>COXI_PART_A1_Hap_5_
TTTAATTTATCAGCATCTTTTTTGG
>COXI_PART_A1_Hap_6_
TCTTATTTATCAACATCTTTTTTGG
>COXI_PART_A1_Hap_7_
TTTAATCTATCAACATCTTTTTTGA
>COXI_PART_A1_Hap_8_
TTTAATTTATCAACATTTATTTTGA
>COXI_PART_A1_Hap_9_
TTTGATTTATCAGCATTTGTTTTGG
>COXI_PART_A1_Hap_10_
TTTGATTTACCAGCATTTGTTTTGG
>COXI_PART_A1_Hap_11_
TTTAATTTATCAACATTTGTTTTGG
>COXI_PART_A1_Hap_12_
TTTAATCTATCAACATTTGTTTTGG
>COXI_PART_A1_Hap_13_
TTTGATTTATCAACATTTGTTTTGG
>COXI_PART_A1_Hap_14_
TTTAATTTATCAGCATTTATTTTGG
>COXI_PART_A1_Hap_15_
TTTGATTTATCAGCATTTATTTTGG
>COXI_PART_A1_Hap_16_
TCTGATTTATCAGCATTTGTTTTGG
>COXI_PART_A1_Hap_17_
TTTAATTTATCAGCATTTGTTTTGG
>COXI_PART_A1_Hap_18_
TCTTATTTATCAGCATCTGTTTTGG
>COXI_PART_A1_Hap_19_
TTTAATTTATCAACATTTGTTTTGA
>COXI_PART_A2_Hap_1_
TTTTTTGGTCATCCTGAGGTTTATA
>COXI_PART_A2_Hap_2_
TTCTTTGGTCATCCTGAAGTATATA
>COXI_PART_A2_Hap_3_
TTCTTTGGTCATCCTGAAGTTTATA
>COXI_PART_A2_Hap_4_
TTTTTTGGACACCCCGAAGTTTATA
>COXI_PART_A2_Hap_5_
TTTTTTGGGCATCCTGAGGTTTATA
>COXI_PART_A2_Hap_6_
TTTTTTGGTCATCCCGAAGTTTATA
>COXI_PART_A2_Hap_7_
TTCTTTGGTCATCCCGAAGTATATA
>COXI_PART_A2_Hap_8_
TTTTTTGGTCATCCAGAGGTTTATA
>COXI_PART_A2_Hap_9_
TTTTTCGGTCATCCAGAAGTTTATA
>COXI_PART_A2_Hap_10_
TTTTTTGGTCATCCAGAAGTTTATA
>COXI_PART_A2_Hap_11_
TTTTTTGGTCATCCGGAAGTTTATA
>COXI_PART_A2_Hap_12_
TTTTTTGGTCATCCTGAAGTTTATA
>COXI_PART_A2_Hap_13_
TTTTTTGGCCATCCAGAAGTTTATA
>COXI_PART_A2_Hap_14_
TTTTTTGGACATCCTGAGGTTTATA
>COXI_PART_A2_Hap_15_
TTCTTTGGCCATCCTGAGGTATATA
>COXI_PART_A2_Hap_16_
TTCTTTGGACATCCTGAAGTATATA
>COXI_PART_A2_Hap_17_
TTTTTTGGTCATCCTGAGGTATATA
>COXI_PART_A2_Hap_18_
TTCTTTGGTCATCCTGAGGTATATA
>COXI_PART_A3_Hap_1_
TTTTAATTCTTCCTGCTTTT
```

>COXI_PART_A3_Hap_2_
TTTTAATTTTGCCTGCTTTT
>COXI_PART_A3_Hap_3_
TTTTGATTCTTCCTGCTTTT
>COXI_PART_A3_Hap_4_
TCCTAATTTTACCTGCTTTT
>COXI_PART_A3_Hap_5_
TTTTGATTTTACCTGCTTTT
>COXI_PART_A3_Hap_6_
TTTTAATTTTGCCTGCATTT
>COXI_PART_A3_Hap_7_
TTTTAATTTTGCCTGCTTTC
>COXI_PART_A3_Hap_8_
TTTTAATTTTACCTGCTTTC
>COXI_PART_A3_Hap_9_
TTTTAATTTTACCTGCTTTT
>COXI_PART_A3_Hap_10_
TTTTAATTCTACCTGCTTTT
>COXI_PART_A3_Hap_11_
TTTTAATTTTACCTGCCTTC
>COXI_PART_A3_Hap_12_
TTTTAATTCTACCTGCTTTC
>COXI_PART_A3_Hap_13_
TCTTAATTCTTCCTGCTTTT
>COXI_PART_B1_Hap_1_
GGTATTATTAGACAAAGTACTCTTT
>COXI_PART_B1_Hap_2_
GGTATTATTAGTCAATCTACTCTTT
>COXI_PART_B1_Hap_3_
GGTATTATCAGTCAGTGTACTTTAT
>COXI_PART_B1_Hap_4_
GGTATTATTAGTCAATGTACTTTAT
>COXI_PART_B1_Hap_5_
GGTATTATTAGTCAATGCACTTTGT
>COXI_PART_B1_Hap_6_
GGTATTATTAGTCAGTGTACTCTTT
>COXI_PART_B1_Hap_7_
GGGATTATTAGTCAAAGTACTTTAT
>COXI_PART_B1_Hap_8_
GGTATTATTAGGCAAAGTACTCTTT
>COXI_PART_B1_Hap_9_
GGTATCATTAGGCAAAGTACTCTTT
>COXI_PART_B1_Hap_10_
GGTATTATTAGTCAAAGTACTCTTT
>COXI_PART_B1_Hap_11_
GGTATTATTAGTCAGTCTACTCTGT
>COXI_PART_B1_Hap_12_
GGTATTATTAGTCAGTGTACTTTAT
>COXI_PART_B1_Hap_13_
GGTATTATTAGTCAAAGTACTTTAT
>COXI_PART_B1_Hap_14_
GGAATTATTAGTCAAAGTACTCTTT
>COXI_PART_B1_Hap_15_
GGTATTATTAGCCAATGTACTTTGT
>COXI_PART_B1_Hap_16_
GGTATTATTAGTCAAAGTACTCTCT
>COXI_PART_B1_Hap_17_
GGTATTATTAGCCAATGTACCTTGT
>COXI_PART_B1_Hap_18_
GGTATTATTAGTCAATGTACCTTGT
>COXI_PART_B1_Hap_19_
GGTATTATTAGTCAATGTACTTTGT
>COXI_PART_B1_Hap_20_
GGTATTATTAGTCAGTGTACTTTGT
>COXI_PART_B1_Hap_21_
GGAATTATTAGACAAAGTACTCTTT
>COXI_PART_B1_Hap_22_
GGAATTATTAGTCAATGTACTTTAT
>COXI_PART_B2_Hap_1_
ACCTTACAGGTAAAAAGGAGGTTTT
>COXI_PART_B2_Hap_2_
ATTTAACTGGTAAGAAAGAGGTTTT
>COXI_PART_B2_Hap_3_
ATTTGACTGGTAAGAAAGAGGTTTT
>COXI_PART_B2_Hap_4_
ATTTAACTGGTAAAAAGGAGGTCTT
>COXI_PART_B2_Hap_5_
ATTTGACTGGTAAGAAAGAAGTTTT
>COXI_PART_B2_Hap_6_
ATTTAACTGGTAAAAAGGAAGTCTT
>COXI_PART_B2_Hap_7_
ATTTGACTGGTAAAAAGGAAGTTTT

>COXI_PART_B2_Hap_8_
ATTTAACTGGTAAAAAGGAAGTTTT
>COXI_PART_B2_Hap_9_
ATTTAACTGGTAAAAAGGAGGTGTT
>COXI_PART_B2_Hap_10_
ATTTAACTGGTAAAAAGGAGGTTTT
>COXI_PART_B2_Hap_11_
ACTTAACTGGTAAAAAGGAGGTGTT
>COXI_PART_B2_Hap_12_
ATTTAACTGGTAAGAAGGAGGTGTT
>COXI_PART_B2_Hap_13_
ATTTGACTGGTAAAAAAGAAGTTTT
>COXI_PART_B2_Hap_14_
ATTTAACCGGTAAAAAAGAGGTATT
>COXI_PART_B2_Hap_15_
ATTTGACTGGTAAGAAGGAGGTGTT
>COXI_PART_B2_Hap_16_
ATTTAACCGGTAAAAAGGAAGTATT
>COXI_PART_B2_Hap_17_
ATTTAACCGGTAAAAAGGAGGTATT
>COXI_PART_B2_Hap_18_
ATTTAACTGGTAAAAAGGAGGTATT
>COXI_PART_B2_Hap_19_
ACTTGACTGGTAAAAAGGAGGTGTT
>COXI_PART_B2_Hap_20_
ATTTAACTGGTAAAAAAGAAGTATT
>COXI_PART_B2_Hap_21_
ATTTAACTGGTAAAAAAGAAGTTTT
>COXI_PART_B2_Hap_22_
ATTTGACTGGTAAAAAGGAGGTGTT
>COXI_PART_B3_Hap_1_
TGGCTCTTTGGGGATAGTTT
>COXI_PART_B3_Hap_2_
TGGTTATTTAGGAATGGTTT
>COXI_PART_B3_Hap_3_
TGGTTATTTAGGGATGGTTT
>COXI_PART_B3_Hap_4_
TGGTTATTTGGGTATGGTTT
>COXI_PART_B3_Hap_5_
TGGTTATTTAGGTATGGTTT
>COXI_PART_B3_Hap_6_
TGGTACTTTAGGTATAATTT
>COXI_PART_B3_Hap_7_
TGGTTACTTAGGAATGGTTT
>COXI_PART_B3_Hap_8_
TGGTTATTTGGGAATGGTTT
>COXI_PART_B3_Hap_9_
TGGTACTTTGGGTATAATTT
>COXI_PART_B3_Hap_10_
TGGTACTTTAGGTATGATTT
>COXI_PART_B3_Hap_11_
TGGTACTTTGGGTATGATTT
>COXI_PART_B3_Hap_12_
CGGTTATTTGGGTATGGTTT
>COXI_PART_B3_Hap_13_
TGGTACCTTGGGTATAATTT
>COXI_PART_B3_Hap_14_
TGGTACCTTAGGTATAATTT
>COXI_PART_B3_Hap_15_
TGGTTCTTTAGGAATGGTTT
>COXI_PART_B3_Hap_16_
TGGTTATCTTGGTATGGTTT
>COXI_PART_B3_Hap_17_
TGGTACTTTAGGTATAATCT
>COXI_PART_B3_Hap_18_
TGGGTATTTGGGTATGGTTT
>COXI_PART_C1_Hap_1_
ATGCTATTTTAAGTATTGGTTTAAT
>COXI_PART_C1_Hap_2_
ATGCTATTTTAAGTATTGGTTTGAT
>COXI_PART_C1_Hap_3_
ACGCTATTTTAAGTATTGGGTTAAT
>COXI_PART_C1_Hap_4_
ACGCTATTCTAAGTATTGGGTTAAT
>COXI_PART_C1_Hap_5_
ACGCTATTTTAAGTATTGGATTAAT
>COXI_PART_C1_Hap_6_
ATGCTATTTTGAGTATTGGTTTAAT
>COXI_PART_C1_Hap_7_
ATGCTATTTTAAGAATTGGTTTAAT
>COXI_PART_C1_Hap_8_
ATGCTATTCTAAGTATTGGTTTAAT

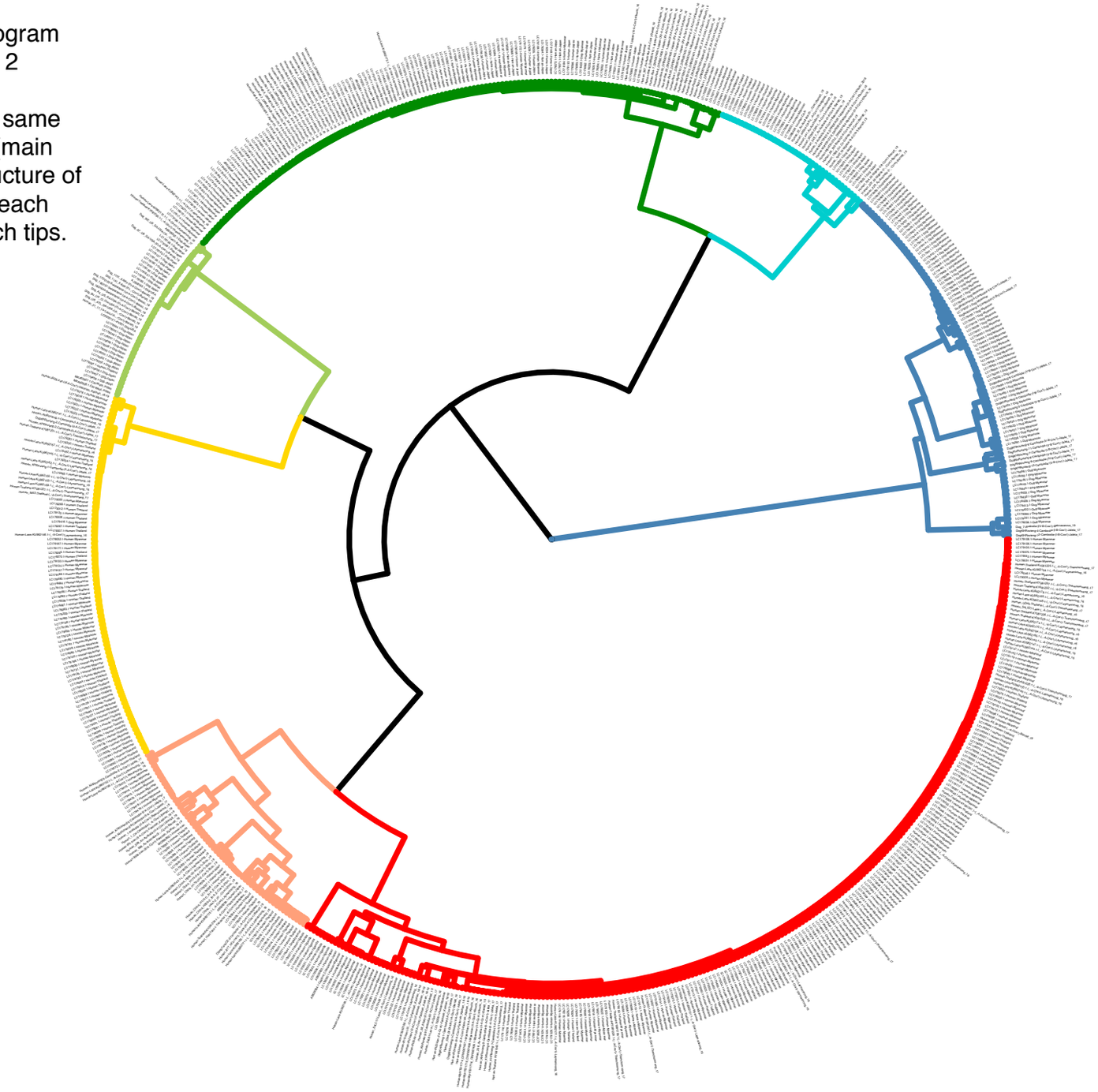
>COXI_PART_C1_Hap_9_
ACGCTATCTTAAGTATTGGATTGAT
>COXI_PART_C1_Hap_10_
ATGCTATTTTAAGTATTGGGTTAAT
>COXI_PART_C1_Hap_11_
ATGCTATTTTAAGTATTGGGTTGAT
>COXI_PART_C1_Hap_12_
ATGCGATTTTAAGTATTGGTTTAAT
>COXI_PART_C1_Hap_13_
ACGCGATTTTAAGTATTGGTTTAAT
>COXI_PART_C1_Hap_14_
ATGCAATTTTAAGTATTGGTTTAAT
>COXI_PART_C1_Hap_15_
ATGCGATTTTAAGTATTGGTTTGAT
>COXI_PART_C1_Hap_16_
ACGCTATTTTAAGTATTGGATTGAT
>COXI_PART_C1_Hap_17_
ATGCTATCTTAAGTATTGGATTGAT
>COXI_PART_C1_Hap_18_
ATGCCATTTTAAGTATTGGTTTGAT
>COXI_PART_C1_Hap_19_
ATGCTATTTTAAGGATTGGTTTAAT
>COXI_PART_C1_Hap_20_
ATGCTATTTTAAGAATTGGTTTGAT
>COXI_PART_C1_Hap_21_
ATGCTATTTTGAGTATTGGTTTGAT
>COXI_PART_C2_Hap_1_
TGGTTGTGTTGTTTGGGCTCATCAT
>COXI_PART_C2_Hap_2_
TGGTTGTGTAGTTTGAGCTCATCAT
>COXI_PART_C2_Hap_3_
TGGTTGTGTAGTTTGAGCTCATCAC
>COXI_PART_C2_Hap_4_
TGGTTGTGTAGTTTGGGCTCATCAC
>COXI_PART_C2_Hap_5_
TGGTTGTGTAGTGTGGGCTCATCAT
>COXI_PART_C2_Hap_6_
TGGTTGTGTAGTATGGGCTCATCAT
>COXI_PART_C2_Hap_7_
TGGTTGTGTAGTTTGAGCTCACCAC
>COXI_PART_C2_Hap_8_
TGGTTGTGTGGTTTGGGCTCATCAT
>COXI_PART_C2_Hap_9_
CGGTTGTGTAGTTTGAGCTCATCAC
>COXI_PART_C2_Hap_10_
TGGTTGTGTGGTTTGAGCTCATCAT
>COXI_PART_C2_Hap_11_
TGGTTGTGTAGTTTGGGCTCATCAT
>COXI_PART_C2_Hap_12_
CGGTTGTGTTGTTTGGGCTCATCAT
>COXI_PART_C2_Hap_13_
TGGTTGTGTTGTTTGAGCTCATCAT
>COXI_PART_C2_Hap_14_
TGGTTGTGTAGTTTGGGCTCACCAC
>COXI_PART_C2_Hap_15_
TGGTTGTGTTGTCTGGGCTCATCAT
>COXI_PART_C2_Hap_16_
TGGTTGTGTTGTTTGAGCACATCAT
>COXI_PART_C3_Hap_1_
ATGTATACTGTTGGTATAGATATTGAT
>COXI_PART_C3_Hap_2_
ATGTATACTGTTGGTATGGATTTTGAT
>COXI_PART_C3_Hap_3_
ATGTATACTGTTGGTATGGATTTCGAT
>COXI_PART_C3_Hap_4_
ATGTATACTGTTGGTATAGATTTTGAT
>COXI_PART_C3_Hap_5_
ATGTATACTGTTGGTATGGATATTGAT
>COXI_PART_C3_Hap_6_
ATGTATACTGTTGGAATGGATTTTGAT
>COXI_PART_C3_Hap_7_
ATGTATACTGTTGGAATGGATTTCGAC
>COXI_PART_C3_Hap_8_
ATGTATACTGTTGGAATGGATTTCGAT
>COXI_PART_C3_Hap_9_
ATATATACTGTTGGTATGGATTTCGAT
>COXI_PART_C3_Hap_10_
ATGTATACTGTTGGAATAGATTTTGAT
>COXI_PART_C3_Hap_11_
ATGTATACCGTTGGTATAGATATTGAT
>COXI_PART_C3_Hap_12_
ATGTATACTGTTGGTATGGACATTGAT

>COXI_PART_C3_Hap_13_
ATGTATACTGTGGGTATAGATTTGGAT
>COXI_PART_C3_Hap_14_
ATGTATACTGTTGGTATATATTTTGAT
>COXI_PART_C3_Hap_15_
ATATATACTGTTGGTATGGATATTGAT
>COXI_PART_C3_Hap_16_
ATGTATACTGTTGGTATGGATTTGGAT
>HVR_I_PART_A_Hap_1_
GTTGGATAACTGAGGTAATTCTTGAGCTAATACACGCTACCTATACCACATTAGTGGTGCGTTTATTTGATTAAACCATTATTTTGGTTGACTCAAAAT
>HVR_I_PART_A_Hap_2_
GTTGGATAACTGAGGTAATTCTTGAGCTAATACACGCTATTTATACCACATTAGTGGTGCGTTTATTTGATTAAACCATTATTTTGGTTGACTCAAAAT
>HVR_I_PART_A_Hap_3_
GTTGGATAACTGACGTAATTCTTGAGCTAATACACGCTATTTATACCACATTAGTGGTGCGTTTATTTGATTAAACCATTATAACGGTTGACTCAAAAT
>HVR_I_PART_A_Hap_4_
GTTGGATAACTGAGGTAATTCTTGAGCTAATACACGCTATTTATACCACATTAGTGGTGCGTTTATTTGATTAAACCATTATAACGGTTGACTCAAAAT
>HVR_I_PART_A_Hap_5_
GTTGGATAACTGAGGTAATTCTTGAGCTAATACACGCTATTTATACCACATTAGTGGTGCGTTTATTTGATTAAACCATTTTAACGGTTGACTCAAAAT
>HVR_I_PART_A_Hap_6_
GTTGGATAACTGAGGTAATTCTTGAGCTAATACACGCTGTTAATACCACATTAGTGGTGCGTTTATTTGATTAAACCATTATATTGGTTGACTCAAAAT
>HVR_I_PART_A_Hap_7_
GTTGGATAACTGAGGTAATTCTTGAGCTAATACACGCTATTTATACCACATTAGTGGTGCGTTTATTTGATTAAACCATTTATACTGGTTGACTCAAAAT
>HVR_I_PART_A_Hap_8_
GTTGGATAACTGAGGTAATTCTTGAGCTAATACACGCTATTTATACCACATTAGTGGTGCGTTTATTTGATTAAACCATTTTATACTGGTTGACTCAAAAT
>HVR_I_PART_A_Hap_9_
GTTGGATAACTGAGGTAATTCTTGAGCTAATACACGCTATTTATACCACATTAGTGGTGCGTTTATTTGATTAAACCATTTTTATATTGGTTGACTCAAAAT
>HVR_I_PART_A_Hap_10_
GTTGGATAACTGAGGTAATTCTTGAGCTAATACACGCTATTTATACCACATTAGTGGTGCGTTTATTTGATTAAACCATTTTATATTGGTTGACTCAAAAT
>HVR_I_PART_A_Hap_11_
GTTGGATAACTGAGGTAATTCTTGAGCTAATACACGCTATTTATACCACATTAGTGGTGCGTTTATTTGATTAAACCATTTTTATATATTGGTTGACTCAAAAT
>HVR_I_PART_A_Hap_12_
GTTGGATAACTGAGGTAATTCTTGAGCTAATACACGCTATTTATACCACATTAGTGGTGCGTTTATTTGATTAAACCATTTATATTGGTTGACTCAAAAT
>HVR_I_PART_B_Hap_1_
ATCCTCGCTGATTTTGTTACTAAAACATACCGTATGTGTATCTGGTTTATCAACTTTCGATGGTAGGGTATTGGCCTACCATGGTTGTGACGGATAACGGAGAATTAGGGT
>HVR_I_PART_B_Hap_2_
ATCCTCGCTGATTTTGTTACAAAAACATACCGTATGTGTATCTGGTTTATCAACTTTCGATGGTAGGGTATTGGCCTACCATGGTTGTGACGGATAACGGAGAATTAGGGT
>HVR_I_PART_B_Hap_3_
ATCCTTGCTGATTTTGTTACTAAAACATACCGTATGTGTATCTGGTTTATCAACTTTCGATGGTAGGGTATTGGCCTACCATGGTTGTGACGGATAACGGAGAATTAGGGT
>HVR_I_PART_C_Hap_1_
TCGACTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGAAAATTACCCAATTTTAGTTTAAAGAGGTAGTGACGAAAAATGACAACCA
>HVR_I_PART_C_Hap_2_
TCGACTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGAAAATTACCCAATTTTAGTTCAAAGAGGTAGTGACGAAAAATGACAACCA
>HVR_I_PART_C_Hap_3_
TCGACTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGAAAATTACCCAATTTTAGTTAAAAGAGGTAGTGACGAAAAATGACAACCA
>HVR_I_PART_C_Hap_4_
TCGACTCCGGAGAGGGAGCCTGAGAAATGGCTACCACATCCAAGGAAGGCAGCAGGCGCGAAAATTACCCAATTTTAGTTAAAAGAGGTAGTGACGAAAAATGACAACCA
>HVR_I_PART_D_Hap_1_
AATATTATTATTAATATTTGGATTGAAAATCTTCAAGTTTAAATAACTTGTTGGTAAAGGAAAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATACCAGC
>HVR_I_PART_D_Hap_2_
AATATTATTATTAATATTTGGATTGAAAATCTTCAAGTTTAAATAACTTGTTGGTAAAGGTTTGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATACCAGC
>HVR_I_PART_D_Hap_3_
AATATTATTATTAATATTTGGATTGAAAATCTTCAAGTATAAATAACTTGTTGGTAAAGGAAAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATACCAGC
>HVR_4_PART_A_Hap_1_
ATTGACAGATTGATAGCTCTTTCATGATTTAGTGGTTGGTGGTGCATGGCCGTTCTTAGTTCGTGGATATGATTTGTCTGGTTGATT
>HVR_4_PART_A_Hap_2_
ATTGACAGAGTGATAGCTCTTTCATGATTTAGTGGTTGGTGGTGCATGGCCGTTCTTAGTTCGTGGATATGATTTGTCTGGTTGATT
>HVR_4_PART_A_Hap_3_
ATTGACAGATTGATAGCTCTTTCATGATCTAGTGGTTGGTGGTGCATGGCCGTTCTTAGTTCGTGGATATGATTTGTCTGGTTGATT
>HVR_4_PART_B_Hap_1_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATAATATTATTTATAAATTTTATATTTTATATAATATTTTATTTTAATAACA
>HVR_4_PART_B_Hap_2_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATAATATTATTTATAAATTTTATATATTATATAATATTTTATTTTAATAACA
>HVR_4_PART_B_Hap_3_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATAATATTATTATTAAATTTTATTTTTATATAATATTTTATTTTAATAACA
>HVR_4_PART_B_Hap_4_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATAATATTATTATTAAATTTTATTTTATATAATATTTTATTTTAATAACA
>HVR_4_PART_B_Hap_5_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATAATATTATTATTAAATTTTATGGTTTTATATAATATTTTATTTTAATAACA
>HVR_4_PART_B_Hap_6_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATAATATTATTGTTAAATTTTATGGTTTTATATAATATTTTATTTTAATAACA
>HVR_4_PART_B_Hap_7_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATAATATTATTGTTAAATTTTATTTTTATATAATATTTTATTTTAATAACA
>HVR_4_PART_B_Hap_9_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATAATATTATTTATAAATTTTATAGTTTTTATATAATATTTTATTTTAATAACA
>HVR_4_PART_B_Hap_8_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATATAATTATTTTGTTTATTTTAATATAAATAATTAATATTTTAATAACA
>HVR_4_PART_B_Hap_10_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATAATATTATTTTAAATTTTATTTAATAATATTTTATTTTAATAACA
>HVR_4_PART_B_Hap_11_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATATTATTATTTTATTATATAAATAATAATTATTTTAATAACA
>HVR_4_PART_B_Hap_12_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATATTATTATTTTTTATATAAATAATAATTATTTTAATAACA

>HVR_4_PART_B_Hap_13_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATATTATTATTTTATTATTTTATATAAATAATTAATATTTTAATAACA
>HVR_4_PART_B_Hap_14_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATATTATTATTTGTTTATTTTTTATATAAATAATTAATATTTTAATAACA
>HVR_4_PART_B_Hap_15_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATATTATTATTTGTTTATTTTATATAAATAATTAATATTTTAATAACA
>HVR_4_PART_B_Hap_16_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATATTATTATTTGTTTATTTTAATATAAATAATTAATATTTTAATAACA
>HVR_4_PART_B_Hap_17_
CCGATAACGAGCGAGACTTTTATGTTATATTAAATATTATTATTTTGTTTATTTTAATATAAATAATTAATATTTTAATAACA
>HVR_4_PART_C_Hap_1_
GATTAATAGTGTTTAACTATTTGAGAGAGAGCAATAACAGGTCTGTGATGCCCTTAGATGTCCGGGGCTGCACGCGCGCTACAAT
>HVR_4_PART_C_Hap_2_
GATTAATAGTGTTTAACTATTTGAGAGAGAGCGATAACGGGTCTGTGATGCCCTTAGATGTCCGGGGCTGCACGCGCGCTACAAT
>HVR_4_PART_C_Hap_3_
GATTAATAGTGTTTAACTATTTGAGAGAGAGCGATAACAGGTCTGTGATGCCCTTAGATGTCCGGGGCTGCACGCGCGCTACAAT
>HVR_4_PART_C_Hap_4_
GATTAATAGTGTTTAACTATTTGAGAGAGAGCGATAACAGGTATGTGATGCCCTTAGATGTCCGGGGCTGCACGCGCGCTACAAT

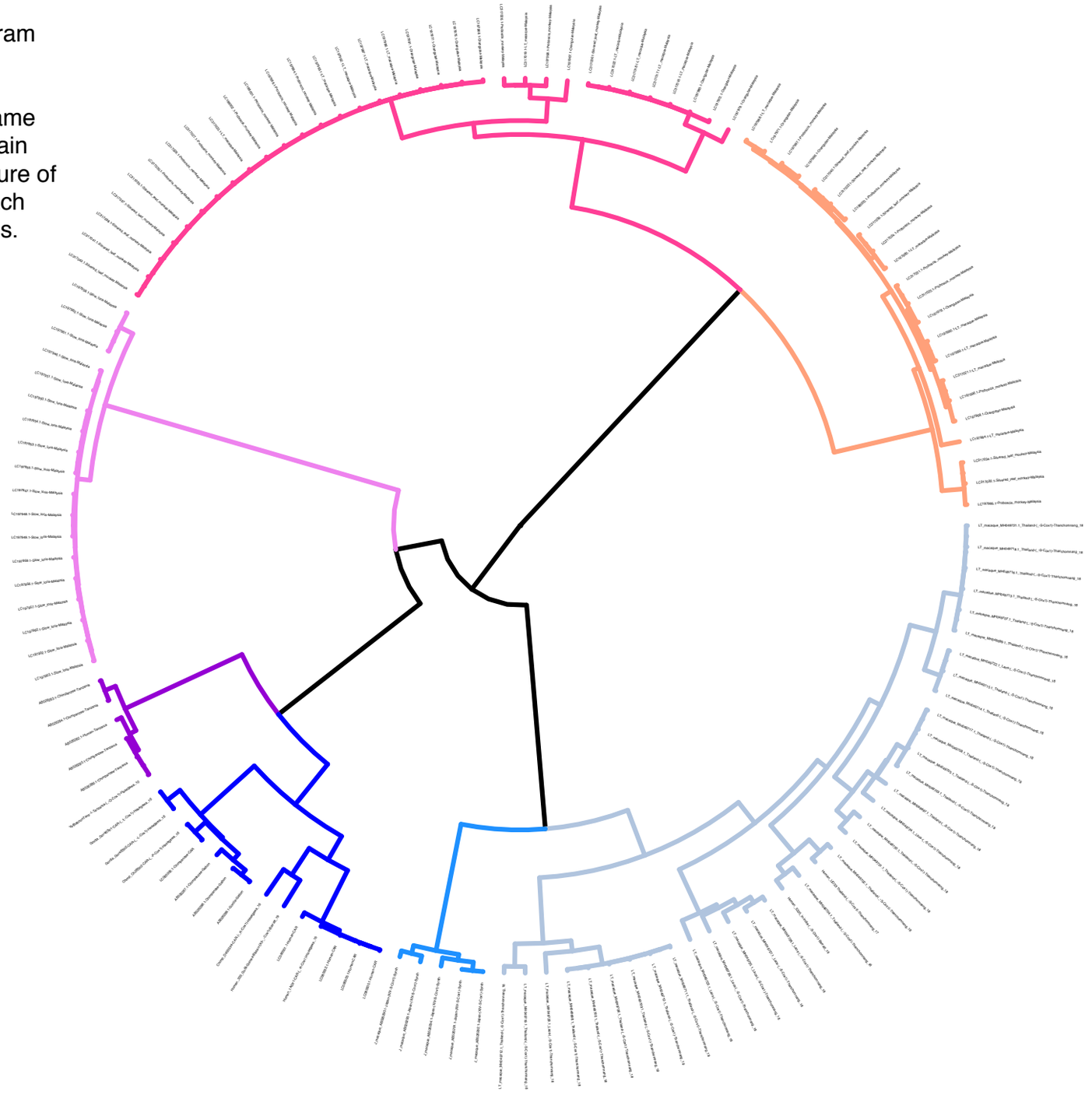
Appendix Part B. Cluster dendrogram identical to that shown in Figure 2

This is a searchable PDF of the same dendrogram shown in Figure 2 (main manuscript text - population structure of S. stercoralis) with the name of each specimen is shown on the branch tips.



Appendix Part C. Cluster dendrogram identical to that shown in Figure 3

This is a searchable PDF of the same dendrogram shown in Figure 3 (main manuscript text - population structure of S. fuelleborni) with the name of each specimen shown on the branch tips.



Appendix Part D. Cluster dendrogram similar to that shown in Figure 3 and Appendix C, though with data generated from Japanese Strongyloides fuelleborni excluded

This searchable PDF shows an alternative analysis of the S. fuelleborni population structure excluding the inferred genotypes from Japanese macaques. This dendrogram demonstrates that exclusion of these specimens does not heavily impact the predicted population structure of S. fuelleborni.