Zoological Studies **65:**46 (2026)

Open Data and Bioinformatics Reveal Species Misidentification in *Chordodes* (Phylum Nematomorpha)

Mattia De Vivo[1,2,*]

[1]*Department of Biogeography, Trier University, Trier, Germany. *Correspondence: E-mail mattiadevivopatalano@gmail.com (De Vivo)*
[2]*Competence Centre for Plant Health, Free University of Bolzano-Bozen, Bolzano/Bozen, Italy*

(Received 21 May 2026 / Accepted 16 July 2026 / Published -- 2026)
Communicated by Jen-Pan Huang

ORCiD
https://orcid.org/0000-0002-9115-1941

The potential usage of genomic open data can help us to understand biodiversity, for example by analyzing species boundaries in morphologically similar species. An example of taxon in which this can be useful is Nematomorpha (horsehair worms), one of the less studied animal phyla, in which species identification can be challenging due to lack of morphological characters. This study started as an evaluation of population-level analyses of effective population size (*Ne*) with open data using the genus *Chordodes*. An RNA sequencing (RNA-seq) dataset originally labelled as *Chordodes fukuii* was first examined by extracting sequences of the mitochondrial barcoding gene cytochrome *c* oxidase subunit I (COXI) from it. After surprising results from such gene, which showed two potential distinct *Chordodes* species in the datasets (*C. formosanus* and *C. japonensis*), further analyses were run by combining these data with a previously published double-digest restriction-site-associated DNA sequencing (ddRADseq) dataset. PCA, adegenet, and ADMIXTURE analyses consistently recovered two distinct genetic clusters, indicating that the RNA-seq dataset contains specimens from two species rather. While most individuals were consistently assigned, some specimens identified as *C. japonensis* based on COXI clustered with individuals labelled as *C. formosanus* or exhibited mixed ancestry when genome-wide markers were analyzed, suggesting possible introgression, incomplete lineage sorting, or biases related to missing data. The study shows how previously released data can be used for evaluating species delimitation, potential previous demographic events, and potential needs in DNA barcoding and genomics for avoiding future misidentification of morphologically similar species.

Keywords: Nematomorph, Taiwan, Japan, Species recognition, Phylogenomics

Citation: De Vivo M. 2026. Open data and bioinformatics reveal species misidentification in *Chordodes* (Phylum Nematomorpha). Zool Stud **65:**46.

BACKGROUND

Recent improvements in genomic sequencing led to the accumulation of data in open repositories, which in turn can be used for analyses and infer new patterns (Plocik and Graveley 2013). Such data can be used to understand biodiversity or help in exploring it. For example, data from open repositories have been used to assemble whole mitogenomes for inferring species relationships (e.g., Vieira and Prosdocimi 2019), analyze potential genomes rearrangements (e.g., Lewin et al. 2024) and infer potential gene loss in lineages, e.g. parasites (e.g., Herlyn et al. 2025).

Among organisms that have seen an increase of available data thanks to sequencing improvements, horsehair worms (Phylum Nematomorpha) are slowly getting more data available after decades of almost complete neglect. Specifically, 13 genomes belonging to 8 different species are available in GenBank (last accessed: June 27, 2026), one of the most popular data repositories for nucleotide data (Clark et al. 2016). This is an increased number compared to the absolute zero that was present before 2023, when three sequenced horsehair worm genomes were made available (Cunha et al. 2023; Eleftheriadi et al. 2024). In addition, Sequencing Read Archives (SRAs) from several projects have been uploaded, for a total of 17 BioProjects with at least a hairworm species sequenced and 104 BioSamples available for species inside this phylum.

So far, the released data showed how horsehair worms lost ortholog metazoan genes (Cunha et al. 2023; Eleftheriadi et al. 2024), which seems to be a common pattern in parasitic Metazoa (Herlyn et al. 2025). Additionally, their mitogenomes have inverted repeats in some protein-coding genes, which are unusual for animals (Mikhailov et al. 2019). Furthermore, transcriptomic data showed potential horizontal gene transfers between freshwater hairworms (class Gordioida) and their definitive hosts, which probably helps the parasites to “manipulate” their hosts (Mishina et al. 2023).

The availability of nucleotide data could also help us to understand horsehair worms’ ecology and taxonomy better. Specifically, although they have a fearsome reputation as “body snatchers” (albeit not all the taxa induce their hosts to jump into water; see Schmidt-Rhaesa 2012 and references therein), Nematomorpha’s popular science notoriety does not translate into a research interest and the phylum is widely regarded as one of the most understudied animal groups (Schmidt-Rhaesa 2012). Given this, we have a relatively limited grasp of their ecology and systematic relationships. For example, the hosts are not known for some species (Schmidt-Rhaesa 2012 and

references therein). Furthermore, horsehair worms have very few morphological characters for taxon identification; because of this, relationships among genera or species in the same genus are very often unclear (Bolek and Hanelt 2024; Schmidt-Rhaesa 2012).

When molecular data were available, it was shown how species complexes or even "genera complexes" are present in Nematomorpha (Anaya et al. 2019; Bolek and Hanelt, 2024; Groß et al. 2025; Hanelt et al. 2015). Therefore, having genomic data can help us understand their phylogeny and if such phylogenetic relationships have also an ecological meaning: for example, Bolek and Hanelt (2024) shows how species in the *Gordius robustus* species complex tend to phylogenetically cluster together according to parasitized final host, stressing the potential use of phylogeny for uncovering ecological interactions.

In this paper, I show how to use a dataset from a previous publication for getting barcodes useful for identifying species and trying to confirm the species ID by combining more data together. For doing so, I extracted cytochrome *c* oxidase subunit I (COXI) sequences from the SRAs of Mishina et al. (2023); such SRAs were labelled as *Chordodes fukuii*, which is one of the horsehair worm species present in Japan (Inoue 1952; Schmidt-Rhaesa 2004) and I wanted to understand their effective population size (*Ne*) dynamics through mitochondrial data together with their host, in a similar fashion as De Vivo et al. (2023a).

BLAST (Altschul et al. 1990) searches with the extracted COXI sequences returned surprisingly high (≈ 100%) BLAST percentage IDs for two different species, and the sequences were very different from the only one sequence available of *C. fukuii* (Table S1). Given this, I changed the focus of the study and tried to understand if there was taxonomic misidentification. To test this hypothesis and assess species boundaries more robustly, I combined the RNA-seq data with a previously published ddRADseq dataset (De Vivo et al. 2023a), which allowed me to have genome-wide data for using population genomic approaches. Specifically, I aimed to determine whether the specimens formed genetically distinct species clusters and whether the original species identifications were supported by genomic data, which allowed me to also check for potential demographic events among the groups. More broadly, this study illustrates how publicly available sequencing datasets can be repurposed for species delimitation and biodiversity research in understudied taxa.

MATERIALS AND METHODS

Used HPC environment and software installation

The workflow followed in this study is summarized in figure 1 and its codes are available in Sup. Data. The reads' download and their assemblies were done by using the High Performance Computing (HPC) at the University of Kaiserslautern-Landau (https://hpc.rz.rptu.de/) in a SLURM environment. I installed the used software and/or the needed dependencies in different Conda environments in the HPC by using Mamba (version 1.5.9; QuantStack 2025).

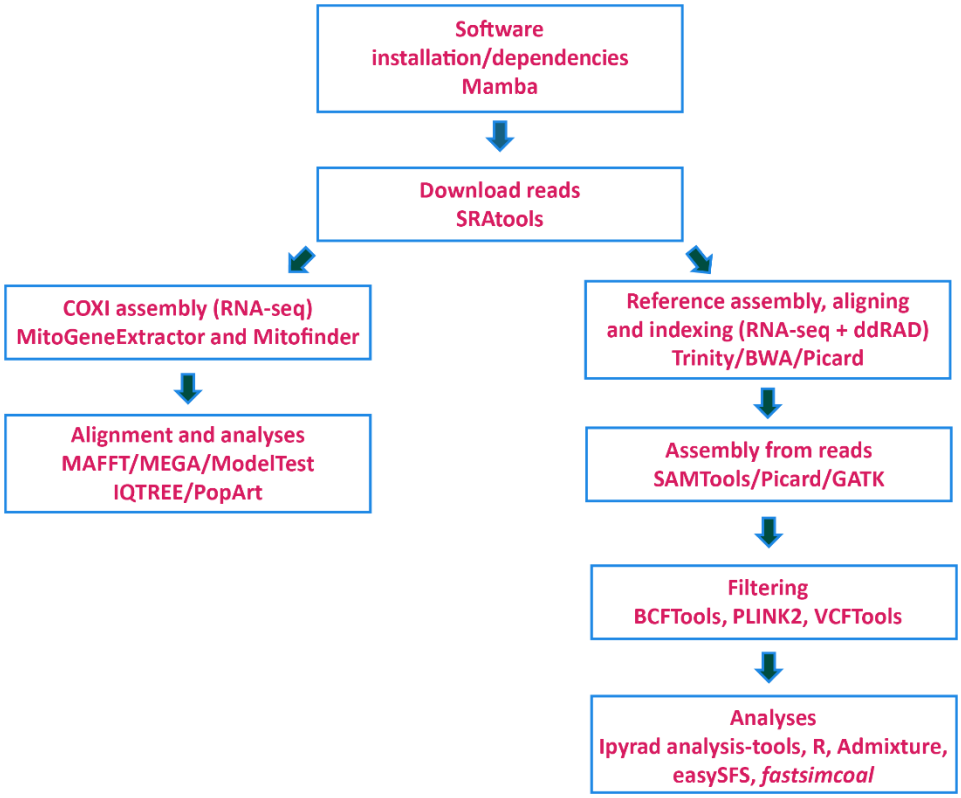

Fig 1. Summary of the bioinformatic workflow followed in this paper.

RNA dataset and initial COXI assembly

Using the SRA Toolkit (version 3.1.1; https://github.com/ncbi/sra-tools), I downloaded the *Chordodes* RNA-seq data from Mishina et al. (2023; BioProject: PRJNA994131), which is made of 19 individual *Chordodes* specimens. Given that the FASTQ files were already trimmed of low-quality reads and adapter sequences, I did not perform any additional trimming.

For assembling protein-coding mitochondrial genes, I used MitoGeneExtractor (version 1.9.5; Brasseur et al. 2023) using the protein annotation of the only *Chordodes* mitogenome available in GenBank at the time of the analysis (accession number: MG257764) as a reference. Given that the SRAs were trimmed, it was not possible to recover all the genes from them (see Brasseur et al. 2023): therefore, I decided to keep only the COXI data for further analyses, given that most of the mitochondrial data of horsehair worms in GenBank belongs to this gene.

For two specimens that presented some "N" in their COXI assembly (SRR25249029 and SRR25249043) and for an additional specimen that did not cluster in the analyses with more genes as it did with COXI (SRR25249039), an additional assembly was attempted by using MitoFinder (version 1.43, Singularity version; Allio et al. 2020) using the GenBank (gb) version of the same mitogenome used for MitoGeneExtractor as reference. A meta-assembly for these specimens' COXI was done by aligning the results from the two software by eye, using MEGA11 (version 11.0.13; Tamura et al. 2021) for visualization and alignment.

BLASting of the COXI sequences, haplotype network and phylogenetic tree

For checking if I was able to get true *Chordodes* sequences from the reads, I used the online version of BLAST (Altschul et al. 1990), with the BLASTn algorithm. All available COXI sequences belonging to the species returned by BLAST searches (Table S1) were downloaded from GenBank (Table S2). I also downloaded sequences from an unidentified *Chordodes* species ("*Chordodes* sp. 3-YT-2023"; LC782468- LC782486; Kuroda et al. 2024) and a sequence labelled as *Chordodes fukuii* from a previous study (AB647244; Sato et al. 2012 and its supplements; Table S2).

After downloading such sequences, I aligned them together with the COXI from the used reference mitogenome and the ones generated from the RNA-seq data using MAFFT (version 7.526; Katoh and Standley 2013) with the L-INS-i algorithm. I manually trimmed such sequences while visualizing them with MEGA 11.

For understanding and visualizing the diversity of the sequences, I generated a haplotype network with the TCS algorithm (version 1.21; Clement et al. 2000) in PopArt (version 1.7; Leigh and Bryant 2015), using the generated alignment. I verified the network's results through a phylogenetic tree: for doing so, I first checked the most likely nucleotide substitution model for the alignment using ModelTest-NG (version 0.1.7; Darriba et al. 2020) implemented in raxmlGUI (version 2.0.13; Edler et al. 2021). I used the most likely model (JC69; Jukes and Cantor 1969) according to the small-sample corrected Akaike Information Criterion (AICc; Sugiura 1978; Sup. Data) for building a phylogeny using IQ-TREE (version 2.2.6; Minh et al. 2020) with 1000 ultrafast bootstrap replicates (UFBoot; Hoang et al. 2018). I visualized the tree with FigTree (version 1.4.4; Rambaut 2018) and rooted it with midpoint rooting.

Assembling of the RNA and ddRAD data and VCF filtering

After getting the phylogenetic tree, I downloaded the 27 ddRAD sequencing data used in the final dataset for *Chordodes formosanus* by De Vivo et al. (2023a; Table S3). This data was uploaded already trimmed too, so no trimming was necessary.

I assembled the RNA-seq and ddRAD reads following a modification of the pipeline followed in Fuhrmann et al. (2023). First, I assembled the transcriptome of the individual SRR25249029, which was the one with most reads from the "after" class (*i.e.*, the worm was extracted from the host and then left for a week in the water; see Mishina et al. 2023 and the GEO dataset of the paper) by using Trinity (version 2.15.2; Haas et al. 2013). Such transcriptome was indexed, and the downloaded reads were aligned to it; both processes were done through BWA (version 0.7.18; Li and Durbin 2010). A following dictionary file of the transcriptome was created with picard (version 3.3.0; Broad Institute 2019).

The SAM files resulting from the alignment were converted into BAM, sorted and indexed by SAMTools (version 1.18; Danecek et al. 2021). The sorted BAM files were initially processed through picard with the "FixMateInformation", "MarkDuplicatesWithMateCigar" and "AddOrReplaceReadGroups" commands and then indexed through SAMTools.

The genotypes were called and outputted in a Variant Call Format (VCF) file through GATK (version 3.8; Van der Auwera et al. 2013), using the "HaplotypeCaller", "BaseRecalibrator", "PrintReads", "CombineGVCFs" and "GenotypeGVCFs" functions. I only kept individuals with at least 5000 retained reads in the final VCF (Sup. Data), which led me to have 27 individuals: 8 from ddRAD and 19 from RNA-seq (Table 1). For checking if the presence of the ddRADseq individuals could cause issues with the analyses, I also made a VCF with only RNA-seq data, by removing individuals in the last two steps of the GATK pipeline (Sup. Data).

I removed indels and filtered for minor allele frequency (MAF) and maximum alleles from the VCF through the "view" function of BCFTools (version 1.21; Danecek et al. 2021), setting MAF at 0.02 and maximum alleles at 2. I then pruned the VCF for linkage disequilibrium (LD) with PLINK2 (version 2.0.0a.6.9; Chang et al. 2015), with the "indep-pairwise" function, using a 50,000 base pairs (bps) window in which LD was checked every 10,000 bps after each step, with a variance inflation factor (VIF) set at 0.5 ("indep-pairwise 50 10 0.5"). Finally, I removed loci with more than 25% missing data from the VCF with VCFTools (version 0.1.16; Danecek et al. 2011), with the "max-missing" function set at 0.75 and kept only variable sites with the "recode" function.

Table 1. Individuals included in the final GATK VCF assembly after filtering

SRA ID	Source	Type of data
SRR22822833	De Vivo et al. (2023a)	ddRAD
SRR22822835	De Vivo et al. (2023a)	ddRAD

SRR22822837	De Vivo et al. (2023a)	ddRAD
SRR22822842	De Vivo et al. (2023a)	ddRAD
SRR22822849	De Vivo et al. (2023a)	ddRAD
SRR22822852	De Vivo et al. (2023a)	ddRAD
SRR22822860	De Vivo et al. (2023a)	ddRAD
SRR22822862	De Vivo et al. (2023a)	ddRAD
SRR25249025	Mishina et al. (2023)	RNA-seq
SRR25249026	Mishina et al. (2023)	RNA-seq
SRR25249027	Mishina et al. (2023)	RNA-seq
SRR25249028	Mishina et al. (2023)	RNA-seq
SRR25249029	Mishina et al. (2023)	RNA-seq
SRR25249030	Mishina et al. (2023)	RNA-seq
SRR25249031	Mishina et al. (2023)	RNA-seq
SRR25249032	Mishina et al. (2023)	RNA-seq
SRR25249033	Mishina et al. (2023)	RNA-seq
SRR25249034	Mishina et al. (2023)	RNA-seq
SRR25249035	Mishina et al. (2023)	RNA-seq
SRR25249036	Mishina et al. (2023)	RNA-seq
SRR25249037	Mishina et al. (2023)	RNA-seq
SRR25249038	Mishina et al. (2023)	RNA-seq
SRR25249039	Mishina et al. (2023)	RNA-seq
SRR25249040	Mishina et al. (2023)	RNA-seq
SRR25249041	Mishina et al. (2023)	RNA-seq
SRR25249042	Mishina et al. (2023)	RNA-seq
SRR25249043	Mishina et al. (2023)	RNA-seq

Population structures' analyses with the VCF file

For analyzing the final datasets and understanding the number of clusters (K) in the VCF, I employed three strategies. The first one used ipyrad (version 0.9.105; Eaton and Overcast 2020) and its analysis-tools suite (https://ipyrad.readthedocs.io/en/latest/API-analysis/index.html). First, the VCF file was converted into a HDF5 database, with the "ld_block_size" argument set as 50,000 (https://ipyrad.readthedocs.io/en/latest/API-analysis/cookbook-vcf2hdf5.html). For the full VCF, I ran a k-means principal component analyses (PCA; https://ipyrad.readthedocs.io/en/latest/API-analysis/cookbook-pca.html), with 100 replicates, 3 populations defined in the "imap" dictionary (Sup. Data), the "minmap" parameter set at 0.5 and the "mincov" parameter set at 0.7. For the RNA-seq only VCF, I defined 2 populations in the "imap" dictionary and used 0.9 as "mincov".

As a second approach, I used functions inside R (version 4.3.1; R Core Team, 2023) implemented in RStudio (version 2023.06.1; Posit team, 2023). I used the package "vcfR" (version 1.15.0; Knaus and Grünwald, 2017) for loading the VCF in R and converting it to the genind format. I then used the snapclust clustering method and the "find.clusters" function implemented

into “adegenet” (version 2.1.10; Beugin et al. 2018; Jombart et al. 2010) for evaluating the ideal number of clusters. In the case of the latter, I used all the available components, and I used the generated data for bar plots.

Finally, I modified the chromosome information in the VCF through BASH scripts to make it compatible with ADMIXTURE (32-bit version, version 1.3.0; Alexander et al. 2009; Sup. Data) and I ran the software, following the scripts available in the manual (https://dalexander.github.io/admixture/admixture-manual.pdf) for Ks ranging from 1 to 3, choosing the best K according to cross validation. All the bar plots (both with data from the “find.clusters” function and the ones from ADMIXTURE) were plotted in R through the “xadmix” package (version 1.0.0; Schönmann, 2022).

For the combined ddRAD-RNA-seq VCF, given that some individuals showed discordant assignments among analyses, I evaluated the effect of missing data on clustering resultsby checking the number of missing sites per individual with the “missing.sort_values” function of the ipyrad analysis-tools, since it is known missing data can affect population assignment (Yi and Latch 2022). I then re-run the three previously described approaches with a VCF without those individuals that had exactly or more than 25% missing data (*i.e.*, SRR25249032, SRR25249035 and SRR25249039; Table S4), which were removed in R through “vcfR” using scripts from Dalapicolla et al. (2021; available at https://github.com/jdalapicolla/LanGen_pipeline_version2/blob/master/PART1/1.1_FILTERING_SNPs.R).)

Demographic events

For understanding if there were hybridization events in the past, I used fastsimcoal (version 2.8.0.0; Excoffier et al. 2021) using the same models made by Meier et al. (2017; see https://speciationgenomics.github.io/fastsimcoal2). Briefly, these models were: "Early gene flow" (i.e., there was hybridization in the past and then it stopped soon after); "No geneflow"; "Recent gene flow" (gene flow in recent times due to secondary contact or similar); "Different gene flow" (different rates of gene flow between old and recent times); "Constant (i.e., ongoing) gene flow".

Compared to the original models, I substituted the 6,000 generations with a potential first hybridization event (TDIV) with a uniform distribution between 1 and 100,000; “1” is the generation time of *Chordodes* (1 year; Chiu 2017), while 100,000 is a time set after checking a population drop according to Stairway Plots in De Vivo et al. (2023a), which may correspond to a population split. However, I preferred not to set that parameter as bound for allowing potential estimates outside of such a boundary. In addition, it is highly likely that the original hybridization event is far more ancient, given the old age of Nematomorpha (Howard et al. 2022; Poinar and

Buckley 2006). Per each model, I set the mutation rate as 2×10^{9} (see De Vivo et al. 2023a for more details).

For running the software, I followed the same methodology as De Vivo et al. (2023b). Briefly, I generated three SFS files using easySFS (version 0.0.1; https://github.com/isaacovercast/easySFS): one with all the outputted individuals, one with only RNA-seq individuals and one with the samples that had less than 25% missing data in the full outputted VCF. Therefore, I had 3 SFS files per model. Although easySFS takes into consideration missing data, I preferred to see if excluding individuals or more loci would have changed the models' results. In the ddRAD-RNA-seq outputs, I selected all the potential *C. japonensis* individuals (defined by COXI, i.e., CJPOT: see the figures in the manuscript), while I chose the number of *C. formosanus* individuals that maximized the number of segregating sites (see the repository's website for more detail). In the case of the RNA-only dataset, I selected all the available individuals.

After generating the SFS files, I ran fastsimcoal with 100 replicates per each model, with 200,000 coalescent simulations (-n) and 50 optimisation cycles (ECM; -L) for estimating parameters based on maximum likelihood (-M). The analyses considered monomorphic sites by pooling together SFS entries with less than 10 SNPs (-C 10) and outputting all SNPs (-s0).

After the models were run, I chose the run with the best estimate for each model with a bash script (available at https://raw.githubusercontent.com/speciationgenomics/scripts/master/fsc-selectbestrun.sh). From such runs, I calculated the Akaike information criterion (AIC; Akaike, 1998) using the highest likelihood values from each model by converting the log10-likelihoods reported by fastsimcoal2 to ln-likelihoods, using the modified script from De Vivo et al. (2023b). I finally calculated AIC weights (Wagenmakers and Farrell 2004) among the two model pairs using the "akaike.weight()" function of the R package 'qpcR' (version 1.4-1; Ritz and Spiess 2008). I also checked the likelihood distributions for model comparison.

From the best model of each SFS, I ran 100 parameter estimation analyses by using 100 bootstrapped VCF files, following scripts from De Vivo et al. (2023b). I then took the values of the time of the events according to the most likely model and made a confidence interval based on mean and standard deviation (SD) through the R "base" package (version 4.3.1; R Core Team, 2023). I made box plots with the "tidyverse" package (version 2.0.0; Wickham et al. 2019) for visualizing the difference in estimated values according to the replicates.

In addition to fastsimcoal, I used triangle plots for evaluating the presence of hybrids and backcrosses in my samples with the R package "triangulaR" (version 0.0.1; Wiens et al. 2025). For doing this, I used the vcfR version of the generated VCF files used also for ADMIXTURE and I

evaluated the results with "1" as allele frequency difference (δ) for finding ancestry-informative markers (AIMs).

RESULTS

COXI results

In total, 200 different sequences were used for the COXI analyses, with 352 bps each (Sup. Data).

The haplotype network shows how the RNA-seq-derived sequences do not cluster together and instead form haplotype groups with previously released *Chordodes formosanus* and *C. japonensis* sequences (Fig. 2). The BLAST results, the output of ModelTest-NG and the haplotype network show how eight COXIs extracted from RNA-seq data (SRR25249025, SRR25249026, SRR25249028, SRR25249030, SRR25249034, SRR25249036, SRR25249037 and SRR25249038) are 100% identical to previously released Japanese *C. formosanus* sequences, while 6 sequences (SRR25249029, SRR25249031, SRR25249032, SRR25249033, SRR25249035 and SRR25249039) are 100% identical to publicly available *C. japonensis* sequences (Table S1; Sup. Data). 5 other extracted sequences (SRR25249027, SRR25249040, SRR25249041, SRR25249042 and SRR25249043) were almost 100% equal to previously released *C. japonensis* sequences (BLASTn similarity percentage ranging between 99.68 and 99.848%; Table S1). The unidentified *Chordodes* specimens from Kuroda et al. (2024), instead, clustered with the only available *C. fukuii* sequence from Sato et al. (2012); said sequences are identical to each other (Fig. 2; Sup. Data).

The phylogenetic tree shows the same conclusions: the sequences extracted from the RNA-seq dataset do not cluster with each other, forming two groups with sequences belonging to *C. formosanus* and *C. japonensis* (bootstrap support=100; Fig. S1; Sup. Data). The top BLAST hit matches the IDs from the tree (Table S1). The unidentified *Chordodes* specimens also cluster here with *C. fukuii* (bootstrap support = 100; Fig. S1; Sup. Data).

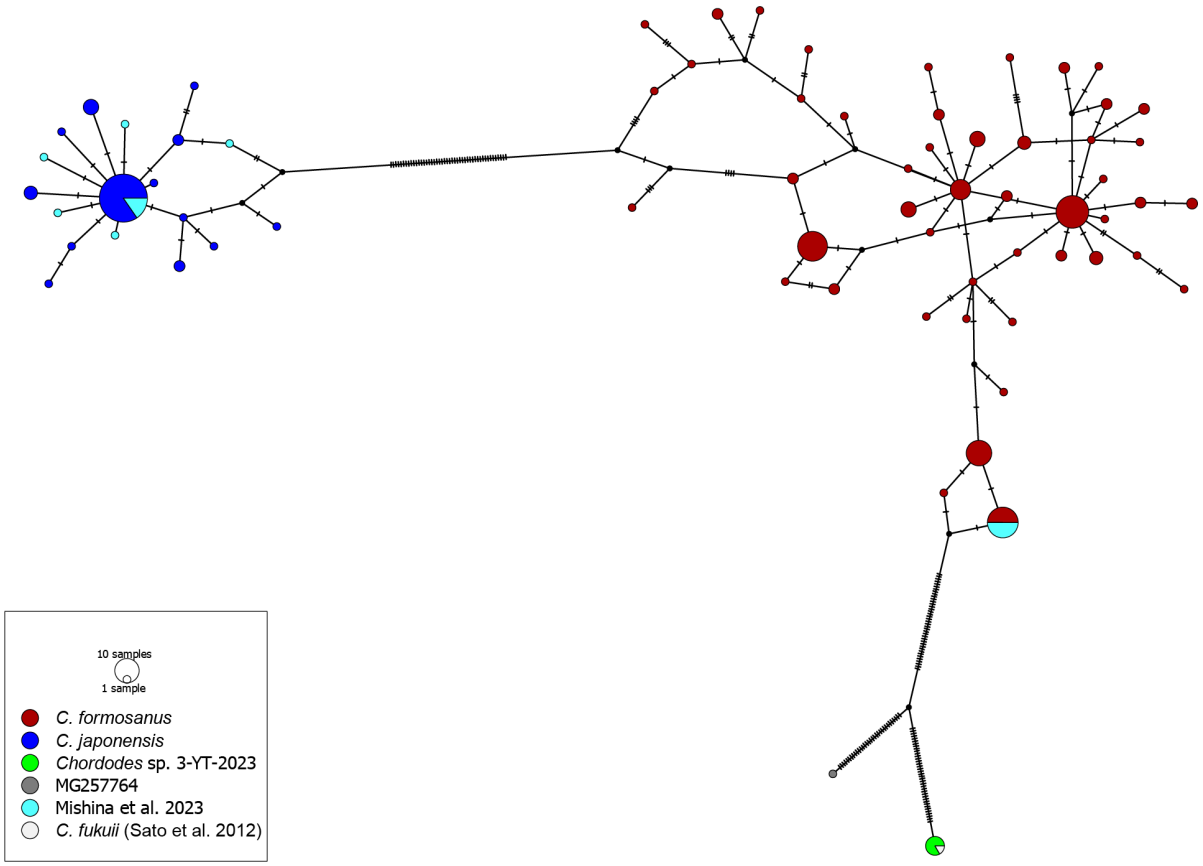

Fig 2. Haplotype network based on the generated 352 bps COXI alignment. The size of the circles represents the number of samples per haplotype. Black circles represent hypothetical haplotypes. The hatchets represent the number of mutations separating each haplotype. The colors represent different groups (see the legend on the bottom left).

ddRAD + RNA-seq clustering analyses

After the filtering, 1,282 variant sites were retained from the 27 final individuals, with 12.92% missing data. In ipyrad analysis-tools, after the "minmap" step, around 1063 SNPs were retained, with 631 of them unlinked (Sup. Data). Using the unlinked SNPs, the PCA shows a split mainly in two groups; one of them is made up of the ddRAD Taiwanese samples and some from RNA-seq, mostly of them labelled as *C. formosanus* according to COXI, although two individuals that were regarded as *C. japonensis* cluster here too. The sequences in the other group were all regarded as *C. japonensis* according to COXI. There was also an individual that formed its own cluster (Fig. 3A).

Such a pattern was also kept after the removal of the three individuals with exactly or more than 25% missing data, with one individual labelled as *C. japonensis* according to COXI still in the hybrid ddRAD/RNA-seq group that should represent *C. formosanus*, although the "rogue single-group" individual was not removed (Fig. 3B).

With the RNA-seq individuals, 32,502 variants sites were retained from 19 individuals, with 8.024% missing data. In ipyrad analysis-tools, after the "minmap" step, around 16,5308 SNPs were

retained, with 7,441 of them unlinked (Sup. Data). A similar pattern to the other two PCAs was found (Fig. 3C).

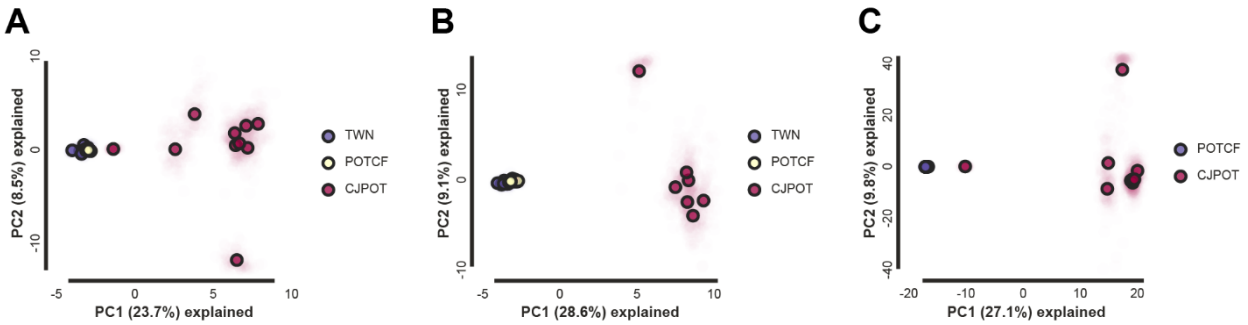

Fig 3. PCAs done with the ipyrad analysis-tools. A, PCA with the full 27 individuals' dataset. B, PCA without individuals with exactly or more than 25% missing data. C, PCA with the individuals from the RNA-seq dataset. "TWN" = *C. formosanus* individuals from the Taiwanese ddRAD dataset. "POTCF" = specimens from Mishina et al. (2023) which were labelled as *C. formosanus* according to COXI. "CJPOT" = specimens from Mishina et al. (2023) which were labelled as *C. japonensis* according to COXI.

Using AIC and snapclust on all the datasets, it was shown how indeed the dataset was highly likely composed of two taxa (Fig. S2). Also here, the population assignment differed compared to the one seen with COXI: specifically, the samples SRR25249029, SRR25249033, SRR25249035 and SRR25249039 were found to be closer to the *C. formosanus*'s cluster (Fig. S3; Sup. Data). SRR25249029, SRR25249039 and SRR25249033 were found to be closer to the *C. formosanus*'s cluster also in the reduced dataset (Fig. S4; Sup. Data). In the RNA-seq-only dataset, SRR25249029 and SRR25249039 still showed a discordant ID compared to COXI (Sup. Fig. 5; Sup. Data).

K = 2 was also the best cross-validation ADMIXTURE result for all the datasets (CV error = 0.54822, 0.56622 and 0.65432 for the full dataset, the ones without individuals with exactly or more than 25% missing data and the RNA-seq only dataset, respectively: they were the lowest CV values for the 3 chosen K values per each dataset; Sup. Data).

ADMIXTURE's results showed that one individual (SRR25249029), which was labelled as labelled as *C. japonensis* according to COXI, clustered 100% with *C. formosanus* in all the generated plots (Fig. 4). Another individual labelled as *C. japonensis* according to COXI, SRR25249039, clustered 100% with *C. formosanus* in the full dataset (Fig. 4A), was removed from the dataset with individuals with less than 25% missing data (Fig. 4B) and clustered 100% with *C. formosanus* again in the RNA-seq only dataset (Fig. 4C).

Two individuals, SRR25249032 and SRR25249035, showed a rough 80 *C. japonensis*:20 *C. formosanus* ancestry ratio in the full datasets (Fig. 4A). They were removed from the dataset with individuals with less than 25% missing data (Fig. 4B). In the RNA-seq-only dataset, SRR25249032 and SRR25249035 still showed more than 80% *C. japonensis* ancestry, mixed with *C. formosanus*

ancestry (≈17% for SRR25249032 and ≈15.5% for SRR25249035; Fig. 4C). SRR25249031 showed an almost complete *C. japonensis* ancestry in the full dataset (Fig. 4A), while it showed roughly 95:5 ancestry ratio in the dataset with fewer individuals (Fig. 4B) and complete *C. japonensis* ancestry in the RNA-seq-only dataset (Fig. 4C). One individual, SRR25249033, showed always 100% *C. japonensis* clustering, minus in the RNA-seq-only derived plot, where it had a 97 *C. japonensis*:3 *C. formosanus* (Fig. 4).

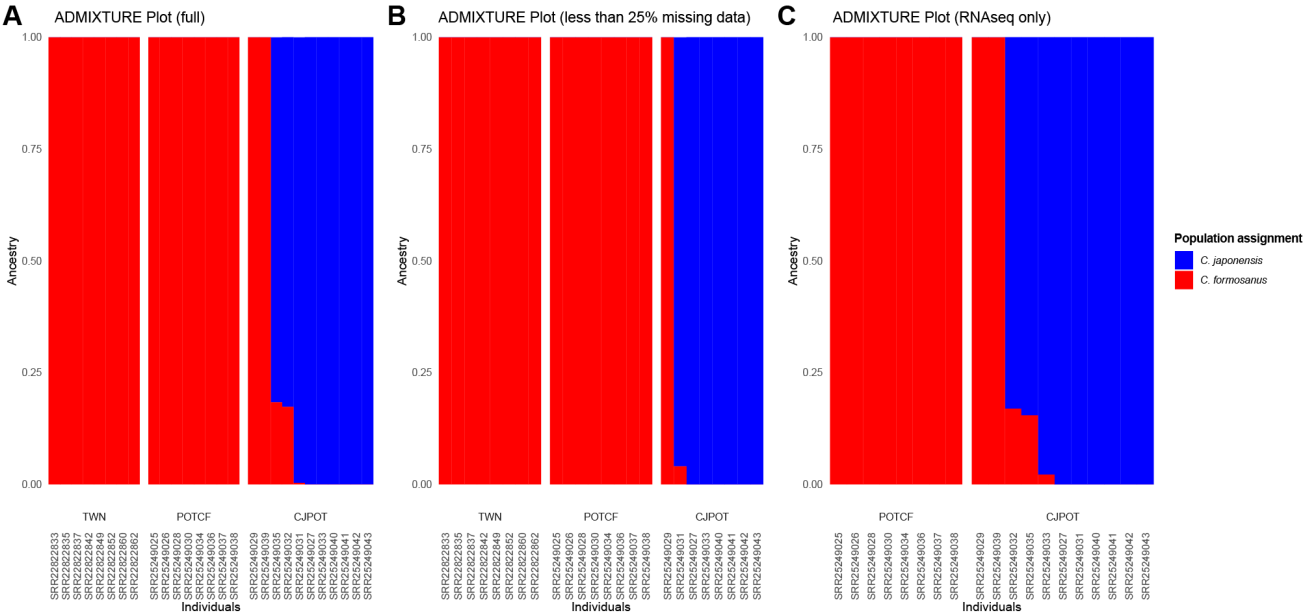

Fig 4. A, ADMIXTURE results with the full 27 individuals' dataset. B, ADMIXTURE results without individuals with exactly or more than 25% missing data. C, ADMIXTURE results with the individuals from the RNA-seq dataset.

Demographic analyses

According to AIC, AIC weights and likelihood distribution, the fastsimcoal model that had different rates of gene flow between split and recent times was the most likely one, with all the datasets (Table 2; Figs. S6-S8; notice how the model with no geneflow has no data for the SFS with all the used individuals, since one deme got extinct in all the 100 simulations with such model).

If both ddRAD and RNA-seq data are used, such demographic events seem to be pretty recent, given that the values of TDIV (first hybridization event) and CHANGM (time of second event) were around 358.81 (mean) ± 84.149 (SD) and 44.13 ± 61.0518 generations ago, respectively, with the SFS with more individuals. Such values were around 759.42 ± 140.571 generations ago for the former and around 94.5 ± 40.554 generations ago for the latter with the SFS excluding the individuals with at least 25% missing data (Fig. 5; Sup. Data).

Table 2. AICs of the fastsimcoal models, rounded up to the second decimal unit. Notice that there is no data for the "no gene flow" model for the full SFS, since a deme got extinct in all the simulations

Model (full SFS)	AIC	ΔAIC	AIC weights
Different gene flow	4800	0	1
Early gene flow	9026.28	4225.88	0
Ongoing gene flow	4835.35	34.945	2.58E-08
Recent gene flow	4891.85	91.44	1.39E-20
Model (less than 25% missing data)	AIC	ΔAIC	AIC weights
Different gene flow	7109	0	1
Early gene flow	12779.96	5671.44	0
No gene flow	13371	6262.42	0
Ongoing gene flow	7216.07	107.55	4.435E-24
Recent gene flow	7204.47	95.95	1.464E-21
Model (RNAseq only)	AIC	ΔAIC	AIC weights
Different gene flow	27586.53	0	1
Early gene flow	27708.30	121.77	3.61E-27
No gene flow	29200.99	1614.46	0
Ongoing gene flow	27697.63	111.1	7.48E-25
Recent gene flow	27697.73	111.2	7.14E-25

With the RNA-seq dataset, these events are postponed of orders of magnitude: specifically, TDIV was around 41,050.9 ± 26,344.67 generations ago and CHANGM was around 1,783.469 ± 965.208 generations ago (Fig. 5; Sup. Data).

After filtering for finding AIMs with δ=1 with "triangulaR", 10 AIMs were left for the full VCF and the one excluding the individuals with at least 25% missing data, while 364 AIMs were left with the RNA-seq VCF. The triangle plots show the presence of at least an individual (SRR25249031) who might be a backcross in the mixed ddRAD-RNA-seq datasets (Fig. 6A-B; Sup. Data). Two individuals (SRR25249032 and SRR25249035) showed a hybrid index > 0 and < 1, showing potential introgression, in the analysis with the full VCF (Fig. 6A; Sup. Data). Three individuals with hybrid index > 0 and < 1, plus heterozygosity > 0 (SRR25249032, SRR25249033 and SRR25249035) are present in the plot using RNA-seq only individuals (Fig. 6C; Sup. Data).

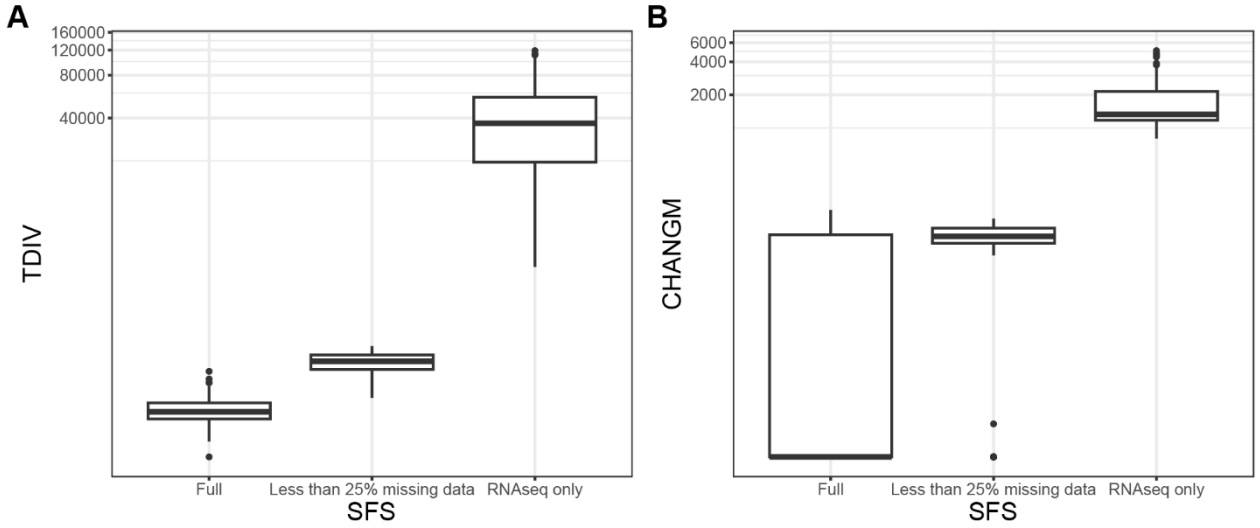

Fig 5. A, Boxplots of the values of TDIV (first hybridization event). B, Boxplots of the values of CHANGM (second hybridization event). All the plots have the y scale converted to log10 ("coord_trans(y = "log10")").

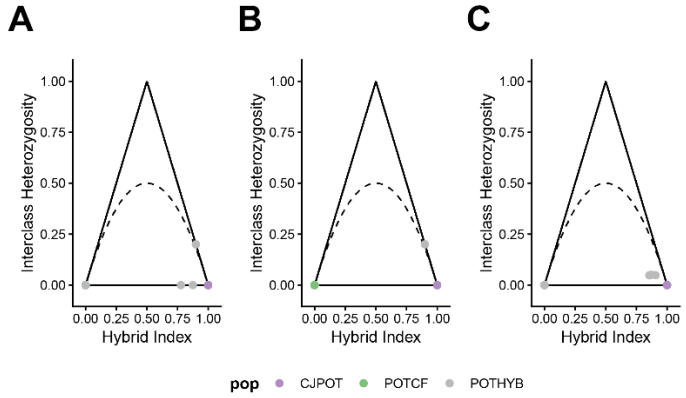

Fig 6. A, Triangle plot for the complete dataset. B, Triangle plot for individuals with less than 25% missing data. C, Triangle plot for the RNAseq dataset. "pop" = population. "CJPOT" = *C. japonensis* individuals according to ADMIXTURE. "POTHYB" = potential hybrids. "POTCF" = *C. formosanus* individuals according to ADMIXTURE.

DISCUSSION

Misidentification from previous publication and the influence of *Chordodes* host usage

In this paper, I report a species misidentification from a previous paper, since what was reported as *Chordodes fukuii* in Mishina et al. (2023) seems to consist of two actual groups. One group corresponds to *C. formosanus*, which is known from Japan and Taiwan (Chiu 2017 and references therein). The other group seems to correspond to *C. japonensis*, which is known from Japan and Korea (Baek 1993; Inoue 1952; Schmidt-Rhaesa 2004), given the high COXI similarity and the clustering with specimens labelled as such using the barcoding gene. Additionally, all the specimens labelled as *C. formosanus* from Mishina et al. (2023) by COXI cluster 100% together in

any analysis, including the ones combing RNA-seq data with ddRADseq from Taiwanese *C. formosanus* specimens.

Generally, species identification in horsehair worms is usually regarded as difficult due to the paucity of morphological characters useful for distinguishing clades, which often are visible only through SEM and/or DIC microscopy (Schmidt-Rhaesa 2012). Chiu et al. (2011) highlighted similarities and dissimilarities between *C. formosanus* and *C. japonensis*, noticing how the former has significantly longer crowned areoles' filaments in female specimens and lacks long-crowned areoles in male individuals. Schmidt-Rhaesa (2004) reports the lack of "characteristic clusters" of areoles in *C. fukuii* compared to *C. japonensis*.

Given that dissimilarities are minimal, some previous work used COXI without SEM for species ID in East Asian *Chordodes*, particularly with damaged specimens (De Vivo et al. 2023a; Kuroda et al. 2024; Tani et al. 2024) or cysts (Chiu et al. 2016). In Chiu et al. (2017), some *Chordodes* specimens were recognized as *C. formosanus* by barcoding after noticing morphological abnormalities, potentially caused by development in the "wrong" definitive host (katydids), which made morphological species identification challenging. De Vivo et al. (2023a) showed how COXI works well for species ID in *C. formosanus*, given the matching ID from such barcoding gene with the one from previous studies (Chiu et al. 2017, 2016, 2011) and the lack of population structure in Taiwanese specimens inferred by ddRAD. Other studies also reported similarities of morphology-based IDs with barcoding based on nuclear and mitochondrial genes in hairworms (Hanelt et al. 2015; Sato et al. 2012; Tobias et al. 2017); the only exception so far is the genus *Gordionus*, which might not be a monophyletic taxon (Anaya et al. 2019; Groß et al. 2025; Sato et al. 2012). The genus *Acutorgordius* might also be another exception, though too few data are present for this taxon for confirming it (Chiu et al. 2020; De Vivo 2024).

It is highly likely the specimens from Mishina et al. (2023) were not correctly processed for morphological taxonomic ID due to the need for high-quality RNA from extraction, since the RNA was extracted from the worms' whole body. This does not change the results of the original paper, since worms clustered according to infection stage in that publication (Mishina et al. 2023). Additionally, the single *C. fukuii* COXI sequence available on GenBank (AB647244; Sato et al. 2012) is not explicitly labelled in such way in the data bank ("*Chordodes* sp. LMB1350000070"), making ID by barcoding for such species very confusing. That being said, the unidentified *Chordodes* species from Kuroda et al. (2024) is highly likely *C. fukuii*, based on both the haplotype network and phylogenetic tree built here. Specifically, sequences from such taxon are extremely similar or equal to the one from Sato et al (2012; similarity ranging from 99.847% to 100% for the full sequences; Sup. Table 5) and even identical in the 352 bp alignment used for building the network and the tree. These examples highlight the need of explicit barcodes for the known

Japanese *Chordodes* species for avoiding species misidentification and highlighting potential undescribed taxa.

A potential reason behind the misidentification could be the collected host: in fact, all the *Chordodes* specimens from Mishina et al. (2023) come from a single host species, the narrow-winged mantis (*Tenodera angustipennis*). Chiu (2017) highlights how no *C. formosanus* specimen was ever reported from any *Tenodera* species. However, reports of *C. formosanus* from *T. angustipennis* in Japan increased recently (Kuroda et al. 2024; Tani et al. 2024), showing how this hairworm is more flexible in host choice than what it was thought. Lack of reports from Taiwanese *T. angustipennis* might be caused by sampling bias or preference for the giant Taiwanese mantis (*Titanodula formosana*, formerly regarded as *Hierodula formosana*; see De Vivo et al. 2023a and references therein), which is the most used host in Taiwan (Chiu 2017 and references therein). Additionally, *C. japonensis* was already reported from *T. angustipennis* in the past (Inoue 1952).

Albeit *C. formosanus* is usually found way more often in giant Asian mantises (subfamily Hierodulinae: Chiu 2017; Chiu et al. 2011; Tani et al. 2024) and *C. japonensis* is usually associated with *Tenodera* species (Chiu et al. 2011; Inoue 1952; Tani et al. 2024), there are reports of both species in different mantises and even katydids (family Tettigoniidae, albeit those might be rarely used and generally not preferred; Chiu 2017; Chiu et al. 2017; Tani et al. 2024). Furthermore, *C. fukuii* also uses *Tenodera* species as definitive hosts (Inoue 1952). Given the following considerations, it is clear that species identification in horsehair worms cannot be based only on checking definitive hosts, albeit host choice might be a phylogenetic signal at genus level.

Mito-nuclear discordance and potential introgression

Some individuals clearly labelled as *C. japonensis* by COXI had some *C. formosanus* ancestry when more than one mitochondrial gene was used, with some of these specimens even showing 100% clustering with *C. formosanus*. However, some of these individuals showed different ancestry ratio according to used dataset. In addition, hybridization in general is not officially reported in hairworms yet.

The presence and the location of missing data might be the reason why ancestry assignation was unclear with different thresholds and data, since it is known that these factors can cause issues in identifying groups with PCA and similar methods (Yi and Latch 2022). This can be particularly true for specimens that had a relative amount of missing data in some analyses, e.g. SRR25249039. Conversely, there were also similar results with different files, methods and thresholds (particularly the ones with PCA and demographic models), highlighting that there might be two past

hybridization events that led to introgression, which is reported in other helminthic taxa (see Detwiler and Criscione 2010).

If such events are confirmed, it is hard to define when they happened. Demographic simulations may have biases when events are recent (Shen et al. 2025) or when they do not account for multiple demographic events (Momigliano et al. 2021). Generally, more data can lead to different and more precise estimates (Shen et al. 2025). Since there was a huge difference in recovered SNPs between the combined ddRAD-RNA datasets and the RNA-seq only dataset, it is highly likely the estimates are more precise with the latter, pushing one of these events back to at least thousands of years ago.

It may be possible that these events happened way farer in the past, given the ancient emergence of freshwater horsehair worms (at least 110/90 millions of years ago; Howard et al. 2022; Poinar and Buckley 2006) and the scarce information on the biogeography of the considered species: specifically, it is possible that their distribution is underestimated. Both species occur and have overlapping distribution in Japan (Chiu et al. 2011; Kuroda et al. 2024; Tani et al. 2024), but *C. japonensis* is also reported from Korea (Baek 1993) and *C. formosanus* is widespread in Taiwan, including the volcanic island of Lyudao (Chiu 2017; De Vivo et al. 2023a). If their geographic origin is not clear, it is challenging to make a potential link between ancient events (*e.g.*, geological ones like the uplift of the island of Taiwan or the formation of the Japanese archipelago) and their population demography, which could help to identify a potential moment where population splits or hybridization could have happened.

If these events and introgression are confirmed, they seem to be biased, given the presence of a mitochondrial lineage from only one species (*C. japonensis*) in these potentially mixed individuals, which might come from male *C. formosanus* breeding with female *C. japonensis*. Females mating with heterospecific males is a known phenomenon (see “Box 1” in Pfennig 2021). If such phenomenon is confirmed, it might be caused by the skewed sex ratio in hairworms (Schmidt-Rhaesa 2012 and references therein) or lower levels of survival in hybrids from male *C. japonensis* and female *C. formosanus*, following the Darwin's Corollary to Haldane's Rule (Turelli and Moyle 2007).

However, another potential valid explanation for the “hybrid” pattern might be incomplete lineage sorting (ILS), caused by ancestral genetic variations that persisted through speciation or, as said earlier, missing data in certain locations. ILS can lead to conflicting topologies, and it is often hard to distinguish it from recent or past introgression (Detwiler and Criscione 2010).

Call for a *Chordodes* genome

All the potential explanations for the mito-nuclear discordance shown here might be proved with data from an outgroup congeneric species and further analyses like ABBA-BABA tests. At the time of the last GenBank check (June 27, 2026), the closest species to *Chordodes* with SRA data on GenBank was highly likely *Parachordes pustulosus* from Europe (Nikolaeva et al. 2023), which however is considered closer based on single gene data and morphology (De Vivo 2024; Nikolaeva et al. 2023; Schmidt-Rhaesa 2012). Furthermore, it may be possible that *Parachordodes* could be phylogenetically far from *Chordodes*. Specifically, although morphology is extremely conserved among clades, Nematomorpha is an ancient group, which probably emerged between mid-Cambrian and mid-Cretaceous (Howard et al. 2022). Freshwater hairworms definitively were already present 110/90 millions of years ago (Howard et al. 2022; Poinar and Buckley 2006); given this, choosing a species from another genera as an outgroup might be risky, given the ancient age of the group, the unresolved phylogenetic questions regarding relationships at different clade levels (family and genus) and the potential ancient split among genera.

Given this, I argue that several questions and doubts could be answered or clarified by sequencing a genome of an outgroup *Chordodes* species and using it as a reference for assembly. Specifically, it has been shown how reference genomes allow to distinguish hybridization from ILS and check for linkage blocks, which can lead to estimation of linkage disequilibrium and better data filtering, while also visualizing potential recombination areas (Theissinger et al. 2023 and references therein). This can be particularly useful for understanding if the results are influenced by the position of missing data (Shen et al. 2025). Furthermore, more loci are recovered through a reference genome: this can help to distinguish recent demographic events from past ones better (Shen et al. 2025). Therefore, a genome assembly of a *Chordodes* species could be important for a better understanding of hairworm ecology and biology.

CONCLUSIONS

I generated COXI data from previously released RNA-seq data and highlighted a case of species misidentification and potential introgression between *Chordodes* species. Furthermore, I also show each needed step for doing these analyses, which can be replicated and used for studies regarding any other species with data available. This paper also shows the need of a reference genome for *Chordodes* species for understanding if there is any current or ancient hybridization in Nematomorpha, one of the most understudied animal groups, and the need in increasing the number of available barcodes for morphologically recognized species.

Acknowledgments: I would like to thank Nico Fuhrmann, whose help and discussion regarding the assembling of the ddRAD-RNAseq dataset was pivotal for making this study. I would also like to thank Vladislav "Uładź" Ivanov for discussion regarding the ddRAD-RNAseq dataset, Jen-Pan Huang for suggestions on how to evaluate potential hybridization, Anzhelika Butenko on comments on a previous draft of the manuscript, plus Wilfried Weber and Hans Recknagel for troubleshooting with the server at Trier University and the anonymous reviewers for suggestions regarding the manuscript.

Authors' contributions: M.D.V. conceived and supervised the study, analyzed the data and wrote the article.

Competing interests: M.D.V. was once affiliated to Academia Sinica, which is the publisher of "Zoological Studies". However, he was not involved in the decision processes regarding the manuscript (*e.g.*, acceptance of refusal).

Availability of data and materials: Data and scripts are available on Zenodo (doi:10.5281/zenodo.21106046) at https://doi.org/10.5281/zenodo.21106046. The scripts are also available on GitHub at: https://github.com/forianfinest/chordodes_species_check. A preprint, based on a previous draft, is available on bioRxiv (doi:10.64898/2026.02.03.703548) at: https://www.biorxiv.org/content/10.64898/2026.02.03.703548v1.

Consent for publication: All the authors consent to the publication of this manuscript.

Ethics approval consent to participate: Since no new specimen was processed for this study, no ethical approval was needed.

REFERENCES

Akaike H. 1998. Information theory and an extension of the maximum likelihood principle. In 'Selected Papers of Hirotugu Akaike', Springer, pp. 199–213. doi:10.1007/978-1-4612-1694-0_15

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res **19:**1655–1664. doi:10.1101/gr.094052.109.

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B et al. 2020. MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Mol Ecol Resour **20:**892–905. doi:10.1111/1755-0998.13160.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology. **215:**403–410. doi:10.1016/S0022-2836(05)80360-2

Anaya C, Schmidt-Rhaesa A, Hanelt B, Bolek MG. 2019. A new species of *Gordius* (Phylum Nematomorpha) from terrestrial habitats in North America. Zootaxa **892:**59–75. doi:10.11646/zootaxa.4927.2.4.

Baek KM. 1993. Two species of genus *Chordodes* (Gordioidea, Nematomorpha) from Korea. Korean J Syst Zool **9:**221–228.

Beugin M-P, Gayet T, Pontier D, Devillard S, Jombart T. 2018. A fast likelihood solution to the genetic clustering problem. Methods Ecol Evol **9:**1006–1016. doi:10.1111/2041-210X.12968.

Bolek MG, Hanelt B. 2024. Chapter 57: Nematomorpha (Phylum): Horsehair Worms [Concepts in Animal Parasitology]. Concepts in Animal Parasitology. doi:10.32873/unl.dc.ciap057.

Brasseur MV, Astrin JJ, Geiger MF, Mayer C. 2023. MitoGeneExtractor: Efficient extraction of mitochondrial genes from next-generation sequencing libraries. Methods Ecol Evol **14:**1017–1024. doi:10.1111/2041-210X.14075.

Broad Institute. 2019. Picard toolkit. Available at: https://broadinstitute.github.io/picard/

Chang CC, Chow CC, Tellier LC et al. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience **4:**s13742-015-0047-8. doi:10.1186/s13742-015-0047-8

Chiu M-C. 2017. Biodiversity of the Taiwanese horsehair worms and the host morphological development manipulated by infection. National Taiwan University Doctoral Dissertation, pp. 1–235. doi:10.6342/NTU201700507.

Chiu M-C, Huang C-G, Wu W-J, Lin Z-H, Chen H-W, Shiao S-F. 2020. A new millipede-parasitizing horsehair worm, *Gordius chiashanus* sp. nov., at medium altitudes in Taiwan (Nematomorpha, Gordiida). ZooKeys **941:**25–48. doi:10.3897/zookeys.941.49100.

Chiu M-C, Huang C-G, Wu W-J, Shiao S-F. 2017. A new orthopteran-parasitizing horsehair worm, *Acutogordius taiwanensis* sp. n., with a redescription of *Chordodes formosanus* and novel host records from Taiwan (Nematomorpha, Gordiida). ZooKeys **683:**1–23. doi:10.3897/zookeys.683.12673.

Chiu M-C, Huang C-G, Wu W-J, Shiao S-F. 2016. Annual survey of horsehair worm cysts in northern Taiwan, with notes on a single seasonal infection peak in chironomid larvae (Diptera: Chironomidae). J Parasitol **102:**319–326. doi:10.1645/15-907.

Chiu M-C, Huang C-G, Wu W-J, Shiao S-F. 2011. A new horsehair worm, *Chordodes formosanus* sp. n. (Nematomorpha, Gordiida) from *Hierodula* mantids of Taiwan and Japan with

redescription of a closely related species, *Chordodes japonensis*. ZooKeys **160:**1–22. doi:10.3897/zookeys.160.2290.

Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2016. GenBank. Nucleic Acids Res **44:**D67–D72. doi:10.1093/nar/gkv1276.

Clement M, Posada D, Crandall KA. 2000. TCS: a computer program to estimate gene genealogies. Mol Ecol **9:**1657–1659. doi:10.1046/j.1365-294x.2000.01020.x.

Cunha TJ, De Medeiros BAS, Lord A, Sørensen MV, Giribet G. 2023. Rampant loss of universal metazoan genes revealed by a chromosome-level genome assembly of the parasitic Nematomorpha. Curr Biol **33:**3514–3521.e4. doi:10.1016/j.cub.2023.07.003.

Dalapicolla J, Prado JR, Percequillo AR, Knowles LL. 2021. Functional connectivity in sympatric spiny rats reflects different dimensions of Amazonian forest-association. J Biogeogr **48:**3196–3209. doi:10.1111/jbi.14281.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E et al. 2011. The variant call format and VCFtools. Bioinformatics. **27:**2156–2158. doi:10.1093/bioinformatics/btr330.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V et al. 2021. Twelve years of SAMtools and BCFtools. Gigascience **10:** giab008. doi:10.1093/gigascience/giab008.

Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B et al. 2020. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. Mol Biol Evol **37:**291–294. doi:10.1093/molbev/msz189.

De Vivo M. 2024. The opportunities from open data: mitogenome assembly from SRAs of horsehair worms (Phylum Nematomorpha). Presented at the 25th Annual meeting of the Society of Biological Systematics (GfBS), Society of Biological Systematics (GfBS), Bonn. doi:10.13140/RG.2.2.31672.72963

De Vivo M, Chen W-Y, Huang J-P. 2023a. Testing the efficacy of different molecular tools for parasite conservation genetics: a case study using horsehair worms (Phylum: Nematomorpha). Parasitology **150:**842–851. doi:10.1017/S0031182023000641.

De Vivo M, Chou M, Wu S, Kuan Y, Chen W et al. 2023b. Genomic tools for comparative conservation genetics among three recently diverged stag beetles (*Lucanus*, Lucanidae). Insect Conserv Divers **16:**853–869. doi:10.1111/icad.12678

Detwiler JT, Criscione CD. 2010. An Infectious Topic in Reticulate Evolution: Introgression and Hybridization in Animal Parasites. Genes **1:**102–123. doi:10.3390/genes1010102.

Eaton DAR, Overcast I. 2020. ipyrad: Interactive assembly and analysis of RADseq datasets. Bioinformatics **36:**2592–2594. doi:10.1093/bioinformatics/btz966.

Edler D, Klein J, Antonelli A, Silvestro D. 2021. raxmlGUI 2.0: A graphical interface and toolkit for phylogenetic analyses using RAxML. Methods Ecol Evol **12:**373–377. doi:10.1111/2041-210X.13512.

Eleftheriadi K, Guiglielmoni N, Salces-Ortiz J, Vargas-Chavez C, Martínez-Redondo GI et al. 2024. The genome sequence of the Montseny horsehair worm, *Gordionus montsenyensis* sp. nov., a key resource to investigate Ecdysozoa evolution. Peer Community Journal **4:**e32. doi:10.24072/pcjournal.381.

Excoffier L, Marchi N, Marques DA, Matthey-Doret R, Gouy A et al. 2021. fastsimcoal2 : demographic inference under complex evolutionary scenarios. Bioinformatics. **37:**4882–4885. doi:10.1093/bioinformatics/btab468.

Fuhrmann N, Prakash C, Kaiser TS. 2023. Polygenic adaptation from standing genetic variation allows rapid ecotype formation. eLife. **12:**e82824. doi:10.7554/eLife.82824.

Groß L, Junker A, Michels M, Plassonke L, Schanné S et al. 2025. Erste Meldungen von *Gordionus violaceus* (Saitenwürmer; Phylum Nematomorpha) für den Raum Trier, mit Kommentaren zur Systematik der Gattung *Gordionus*. Dendrocopos **51:**103–110.

Haas BJ et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols **8:**1494–1512. doi:10.1038/nprot.2013.084.

Hanelt B, Schmidt-Rhaesa A, Bolek MG. 2015. Cryptic species of hairworm parasites revealed by molecular data and crowdsourcing of specimen collections. Mol Phylogenet Evol **82:**211–218. doi:10.1016/j.ympev.2014.09.010.

Herlyn H, Hembrom AA, Tosar J-P, Mauer KM, Schmidt H et al. 2025. Substantial hierarchical reductions of genetic and morphological traits in the evolution of rotiferan parasites. Genome Biol Evol **7:**evaf124. doi:10.1093/gbe/evaf124.

Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol **35:**518–522. doi:10.1093/molbev/msx281.

Howard RJ, Giacomelli M, Lozano-Fernandez J, Edgecombe GD, Fleming JF et al. 2022. The Ediacaran origin of Ecdysozoa: integrating fossil and phylogenomic data. J Geol Soc **179:**jgs2021-107. doi:10.1144/jgs2021-107.

Inoue I. 1952. カマキリ(*Tenodera*)に寄生するハリガネムシ類(Gordiacea)の分類 [Classification of Japanese *Chordodes* (Gordiacea)]. Zool Mag, Tokyo **61:**27–32. doi:10.34435/zm003364.

Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics **11:**94. doi:10.1186/1471-2156-11-94.

Jukes TH, Cantor CR. 1969. Evolution of Protein Molecules. In 'Mammalian Protein Metabolism', Elsevier, pp. 21–132. doi:10.1016/B978-1-4832-3211-9.50009-7

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **30:**772–780. doi:10.1093/molbev/mst010.

Knaus BJ, Grünwald NJ. 2017. vcfR : a package to manipulate and visualize variant call format data in R. Mol Ecol Res **17:**44–53. doi:10.1111/1755-0998.12549.

Kuroda K, Kuroda T, Nishino H, Takami Y. 2024. Sex-specific manipulation of sexually cannibalistic mantid mating behavior by hairworms. Behav Ecol **35:**arae071. doi:10.1093/beheco/arae071.

Leigh JW, Bryant D. 2015. popart : full-feature software for haplotype network construction. Methods Ecol Evol **6:**1110–1116. doi:10.1111/2041-210X.12410.

Lewin TD, Liao IJ-Y, Luo Y-J. 2024. Annelid Comparative Genomics and the Evolution of Massive Lineage-Specific Genome Rearrangement in Bilaterians. Mol Biol Evol **41:**msae172. doi:10.1093/molbev/msae172.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics **26:**589–595. doi:10.1093/bioinformatics/btp324.

Meier JI, Sousa VC, Marques DA, Selz OM, Wagner CE et al. 2017. Demographic modelling with whole-genome data reveals parallel origin of similar *Pundamilia* cichlid species after hybridization. Mol Ecol **26:**123–141. doi:10.1111/mec.13838.

Mikhailov KV, Efeykin BD, Panchin AY, Knorre DA, Logacheva MD et al. 2019. Coding palindromes in mitochondrial genes of Nematomorpha. Nucleic Acids Res **47:**6858–6870. doi:10.1093/nar/gkz517.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD et al. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol **37:**1530–1534. doi:10.1093/molbev/msaa015.

Mishina T, Chiu M-C, Hashiguchi Y, Oishi S, Sasaki A et al. 2023. Massive horizontal gene transfer and the evolution of nematomorph-driven behavioral manipulation of mantids. Curr Biol **33:**4988-4994.e5. doi:10.1016/j.cub.2023.09.052.

Momigliano P, Florin A-B, Merilä J. 2021. Biases in Demographic Modeling Affect Our Understanding of Recent Divergence. Mol Biol Evol **38:**2967–2985. doi:10.1093/molbev/msab047.

Nikolaeva O, Beregova A, Efeykin B, Miroliubova T, Zhuravlev A et al. 2023. Expression of Hairpin-Enriched Mitochondrial DNA in Two Hairworm Species (Nematomorpha). Int J Mol Sci **24:**11411. doi:10.3390/ijms241411411.

Pfennig KS. 2021. Biased Hybridization and Its Impact on Adaptive Introgression. Trends Eco Evol **36:**488–497. doi:10.1016/j.tree.2021.02.010.

Plocik AM, Graveley BR. 2013. New Insights from Existing Sequence Data: Generating Breakthroughs without a Pipette. Mol Cell **49**:605–617. doi:10.1016/j.molcel.2013.01.031.

Poinar G, Buckley R. 2006. Nematode (Nematoda: Mermithidae) and hairworm (Nematomorpha: Chordodidae) parasites in Early Cretaceous amber. J Invertebr Pathol **93:**36–41. doi:10.1016/j.jip.2006.04.006.

Posit team. 2023. RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA

QuantStack. 2025. mamba. Available at: https://github.com/mamba-org/mamba

R Core Team. 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Rambaut A. 2018. FigTree, version 1.4.4. Available at: http://tree.bio.ed.ac.uk/software/figtree/

Ritz C, Spiess A-N. 2008. qpcR: an R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis. Bioinformatics **24:**1549–1551. doi:10.1093/bioinformatics/btn227.

Sato T, Watanabe K, Tamotsu S, Ichikawa A, Schmidt-Rhaesa A. 2012. Diversity of nematomorph and cohabiting nematode parasites in riparian ecosystems around the Kii Peninsula, Japan. Can J Zool **90:**829–838. doi:10.1139/z2012-048.

Schmidt-Rhaesa A. 2012. Nematomorpha. In 'Schmidt-Rhaesa', Handbook of Zoology. Gastrotricha, Cycloneuralia and Gnathifera 1, De Gruyter, Berlin, pp. 29–145. doi:10.1515/9783110272536.29.

Schmidt-Rhaesa A. 2004. Japanese Horsehair Worms (Nematomorpha) from the Lake Biwa Museum. Species Divers **9:**97–107. doi:10.12782/specdiv.9.97.

Schönmann L. 2022. xadmix: Subsetting and Plotting Optimized for Admixture Data. doi:10.32614/CRAN.package.xadmix.

Shen Z-Y, Le M-H, Chou M-H, Huang J-P. 2025. A reference genome enhances the power to detect signatures of recent anthropogenic impact in genomic data: a lesson learned from a stag beetle system. BMC Biology **23:**205. doi:10.1186/s12915-025-02307-7.

Sugiura N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections: Further analysis of the data by akaike's. Communications in Statistics - Theory and Methods **7:**13–26. doi:10.1080/03610927808827599.

Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol **38:**3022–3027. doi:10.1093/molbev/msab120.

Tani S, Tettey PA, Maruta R, Kodama A, Saito H et al. 2024. Host range differences between two species of freshwater horsehair worm (Nematomorpha: Chordiidae) *Chordodes japonensis* and *C. formosanus* in Japan. Parasitol Int **99:**102847. doi:10.1016/j.parint.2023.102847.

Theissinger K et al. 2023. How genomics can help biodiversity conservation. Trends in Genetics. **39:**545–559. doi:10.1016/j.tig.2023.01.005.

Tobias ZJC, Yadav AK, Schmidt-Rhaesa A, Poulin R. 2017. Intra- and interspecific genetic diversity of New Zealand hairworms (Nematomorpha). Parasitology **144:**1026–1040. doi:10.1017/S0031182017000233.

Turelli M, Moyle LC. 2007. Asymmetric Postmating Isolation: Darwin's Corollary to Haldane's Rule. Genetics **176:**1059–1088. doi:10.1534/genetics.106.065979.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G et al. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr Protoc Bioinf **43:**11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43.

Vieira GA, Prosdocimi F. 2019. Accessible molecular phylogenomics at no cost: obtaining 14 new mitogenomes for the ant subfamily Pseudomyrmecinae from public data. PeerJ **7:**e6271. doi:10.7717/peerj.6271.

Wagenmakers E-J, Farrell S. 2004. AIC model selection using Akaike weights. Psychon Bull Rev **11:**192–196. doi:10.3758/BF03206482.

Wickham H et al. 2019. Welcome to the tidyverse. Journal of Open Source Software **4:**1686. doi:10.21105/joss.01686.

Wiens BJ, DeCicco LH, Colella JP. 2025. triangulaR: an R package for identifying AIMs and building triangle plots using SNP data from hybrid zones. Heredity **134:**251–262. doi:10.1038/s41437-025-00760-2.

Yi X, Latch EK. 2022. Nonrandom missing data can bias Principal Component Analysis inference of population genetic structure. Mol Ecol Resour **22:**602–611. doi:10.1111/1755-0998.13498.

Supplementary materials

Table S1. BLASTn results for the COXI extracted from RNA SRAs of Mishina et al. (2023), with also a similarity comparison with the only available *C. fukuii* sequence from Genbank (Sato et al. 2012).

Table S2. GenBank accession numbers of the downloaded COXI sequences.

Table S3. SRAs of all the downloaded ddRADseq data.

Table S4. The amount of missing data in the final VCF assembly according to ipyrad analysis-tools.

Table S5. Similarities between the COXI of the unidentified *Chordodes* specimens from Kuroda et al. (2024) with the only available *C. fukuii* sequence from Genbank (Sato et al. 2012).

Fig. S1. Collapsed phylogenetic tree based on COXI sequences, with also the SRA access number of the sequences from Mishina et al. (2023) and where they cluster. The unprocessed tree is available in Supplementary Data.

Fig. S2. AIC calculated by snapclust. A, AIC for the whole dataset. B, AIC for the individuals with less than 25% missing data. C, AIC for the RNAseq individuals.

Fig. S3. Composition plot from “find.clusters”, with the full dataset.

Fig. S4. Composition plot from “find.clusters”, with individuals with less than 25% missing data.

Fig. S5. Composition plot from “find.clusters”, with the RNA-seq individuals.

Fig. S6. Likelihood distribution for the *fastsimcoal* models with the full SFS. Notice that there is no data for the “no gene flow” model, since a deme got extinct in all the simulations.

Fig. S7. Likelihood distribution for the *fastsimcoal* models with the SFS without the individuals with more than 25% missing data.

Fig. S8. Likelihood distribution for the *fastsimcoal* models with the SFS with only the RNAseq individuals.