Unravelling phylogenetic relationships within the genus *Lispe* (Diptera: Muscidae) through genome-assisted and *de novo* analyses of RAD-seq data

Kinga Walczak[a], Marcin Piwczyński[a], Thomas Pape[b], Nikolas P. Johnston[c,d], James F. Wallman[d], Krzysztof Szpila[a], Andrzej Grzywacz[a*]

[a]Department of Ecology and Biogeography, Faculty of Biological and Veterinary Sciences, Nicolaus Copernicus University in Toruń, Toruń, Poland
[b]Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
[c]Molecular Horizons, School of Chemistry and Molecular Bioscience, University of Wollongong, Wollongong, New South Wales, Australia
[d]Faculty of Science, University of Technology Sydney, Ultimo, New South Wales, Australia

*Corresponding authors: Kinga Walczak (KW), Andrzej Grzywacz (AG), Department of Ecology and Biogeography, Nicolaus Copernicus University in Toruń, Lwowska 1, 87-100 Toruń, Poland, Email: KW: walczak.kinga00@gmail.com; AG: hydrotaea@gmail.com.

Keywords: *de novo* assembly, nanopore sequencing, genome, phylogeny, reference-based assembly

Abstract

Lispe represents a species-rich genus within the family Muscidae. The current subdivision of *Lispe* species into species groups is based mainly on adult morphology and ecology, with the only available phylogenetic study based on three molecular markers. Nonetheless, certain species groups remain unclear, and the relationships and composition of these groups are still unresolved. This study employs restriction-site associated DNA sequencing (RAD-seq) with both reference-based and *de novo* reads assembly approaches to investigate relationships within *Lispe*. To apply a reference-based approach we utilised Oxford Nanopore Technologies long read sequencing to assemble a draft genome of *L. tentaculata*. The resulting topologies of phylogenetic trees are well-supported and relatively consistent, here divided into three main clades. One comprises the *palposa*-, *rigida*- and *caesia*-groups; another includes the *nicobarensis*-, *nivalis*-, *scalaris*- and *tentaculata*-groups; and the next consists of the *longicollis*-, *desjardinsii*-, *uliginosa*- and *kowarzi*-groups. The primary discrepancy between topologies obtained under various analytical approaches is the relationship between the *leucospila*-group and all other ingroup taxa, being a sister taxon either to all remaining *Lispe* or to a clade consisting of the *longicollis*-, *desjardinsii*-, *uliginosa*- and *kowarzi*-groups. *Lispe polonaise*, included for the first time in a molecular phylogenetic analysis, is nested within the *caesia*-group. Similarly, *L. capensis* and the hitherto unassigned *L. mirabilis* belong to the *tentaculata*-group. Our study confirms the validity of the 14 species groups currently recognised in the genus *Lispe*.

1. Introduction

Lispe Latreille, 1796 is a genus of Diptera widespread worldwide, except New Zealand (Hennig, 1965; Vikhrev, 2015). The genus contains 163 described species (Pont, 2019), but extensive taxonomic changes and description of new species is ongoing recently (Fig. 1) (Vikhrev, 2014, 2016, 2020, 2021; Zielke, 2018; Pont, 2019). Species of *Lispe* can be easily distinguished from other muscid genera by the combination of an apically dilated palpus and a setulose anepimeron (Curran, 1937; Snyder, 1954; Pont, 2019). Adults are mainly observed on river banks, along the shores of lakes, seas and oceans or in other wetlands areas (Snyder, 1954; Vikhrev, 2011a), and are commonly known as predators of small insects, e.g., Diptera (Pont, 2019) and Coleoptera (Steidle *et al.*, 1995). Their most common prey are other flies of families such as Psychodidae and Milichiidae (Williams, 1938), Muscidae (Hennig, 1960), Chironomidae and Culicidae (Snyder, 1954; Shinonaga & Kano, 1983; Pont, 2019). Due to their predation of culicid and simuliid populations, species of *Lispe* serve as biological control agents and thus can be considered of economic importance (Snyder, 1954; van Emden, 1965; Werner *et al.*, 2014). However, representatives of *Lispe* exhibit various feeding strategies (Vikhrev, 2011a). *Lispe binotata* Becker, 1914 feeds on invertebrate carrion (Vikhrev, 2011a), and active hunting was observed for example in *L. geniseta* Stein, 1909, *L. tentaculata* (De Geer, 1776) and *L. pygmaea* (Fallén, 1825) (Vikhrev, 2011a; Werner *et al.*, 2014). *Lispe caesia* Meigen, 1826 successfully attacks much larger flies (Hennig, 1960) and *L. candicans* Kowarz, 1892 can penetrate the cuticle of adult beetles (Steidle *et al.*, 1995). In some species of *Lispe*, males perform spectacular courtship dances around passive females, making them valuable models for studying mating behaviour (Frantsevich & Gorb, 2006; Butterworth & Wallman, 2022). Larvae of *Lispe* are obligatory carnivores (Skidmore, 1985), which develop in wet sand or mud with high organic content (Séguy, 1937; Hennig, 1960; Skidmore, 1985), where they feed on aquatic invertebrates (Pont, 2019).

Lispe forms a monophyletic group supported by several apomorphies listed by Hennig (1960, 1965) and recently confirmed by a molecular study (Gao *et al.*, 2022). The systematic position of *Lispe* within Muscidae has changed over the years. The genus has been placed in its own subfamily Lispinae (Malloch, 1923; Séguy, 1937; van Emden, 1941, 1965; Snyder, 1949) or in the tribe Limnophorini of Mydaeinae (Karl, 1928), until Hennig (1960) classified *Lispe* in the Limnophorini of Coenosiinae based on morphological data from eggs, larvae and adults. The classification proposed by Hennig has been adopted by the majority of later authors (Skidmore, 1985; Werner *et al.*, 2014; Pont, 2019), and the relationship between *Lispe* and *Limnophora* Robineau-Desvoidy, 1830 has recently been confirmed by molecular studies

(Kutty *et al.*, 2010; Ge *et al.*, 2016; Grzywacz *et al.*, 2021; Gao *et al.*, 2022). The most recent hypotheses consider *Lispe* in Lispini of Coenosiinae (Fan, 2008), in Limnophorini of Coenosiinae (Pont, 1986; Grzywacz *et al.*, 2021) or in Coenosiinae (Haseyama *et al.*, 2015). Attempts to organise species of *Lispe* into smaller units have been a matter of debate for many years. Currently, the genus is divided into several species groups defined by leg and body chaetotaxy, characters of the male terminalia, as well as the ecology of adults (Hennig, 1960; Pont, 2019). Even though Snyder (1954) is believed to be the first to have proposed such a subdivision, van Emden (1941) tentatively used the term "*tentaculata*-group" and Paterson (1953) used '*Lispe leucospila*-group'. To date, *Lispe* has been classified into 14 species groups (Supp. Table S1). Snyder (1954) proposed three species groups for Nearctic *Lispe,* i.e., the *tentaculata* species group, the *uliginosa* species group and the *palposa* species group. Hennig (1960) added additional Palaearctic *Lispe* to Snyder's species groups and separated another three groups: the *scalaris* species group, the *caesia* species group and the *longicollis* species group, with the latter divided into two subgroups, and he also left several Palaearctic species unassigned. Most recently, the Vikhrev has contributed significantly to the ordering of *Lispe* by proposing and defining further species groups and revising the existing ones, such as: *leucospila* (Vikhrev 2011b), *nivalis* and *rigida* (Vikhrev 2012b), *desjardinsii*, *kowarzi* and *nana* (Vikhrev 2014), *nicobarensis* (Vikhrev 2015) and *ambigua*, *dichaeta*, *geniseta*, *pumila* and *pygmaea* (Vikhrev 2016). Some groups are well-defined, like the *nivalis*-group (Vikhrev, 2012a) and the *palposa*-group (Vikhrev, 2015), while others remain unclear, e.g., the *caesia*-group (Vikhrev *et al.*, 2016). Moreover, five species complexes have been proposed based on similarities in ecology rather than morphology (Vikhrev, 2016). Previous authors did not conduct formal morphology-based phylogenetic analyses, yet hypotheses based on morphological evidence suggest that the *nivalis*-group is closely related to the *tentaculata*-group (Vikhrev, 2012a), and that the *nana*-complex has an intermediate position between the *tentaculata*-group and the *scalaris*-group (Vikhrev, 2014). These four groups form the *tentaculata* supergroup, which is additionally supported by their similar ecological association with freshwater habitats (Vikhrev, 2014). Furthermore, Vikhrev (2014) stated that species of the *desjardinsii*-group resemble those of the *longicollis*-group, and that the *palposa*-group appears to be closely related to the *rigida*-group. A recent molecular study based on three genes investigated the limits of some species groups and confirmed some of those phylogenetic hypotheses (Gao *et al.*, 2022). However, tree topologies were affected by different inference methods and the relationships between species groups emerged with low to moderate nodal support values, especially along the backbone of the phylogenetic tree.

The rapidly decreasing costs of next-generation sequencing (NGS) make it feasible to obtain genomic-scale data in a relatively short time (Metzker, 2010). Advances in high-throughput sequencing approaches over the past decade have proven to be advantageous for systematics and population genetics studies (Hohenlohe *et al.*, 2010; Rubin *et al.*, 2012). Restriction site-associated DNA sequencing (RAD-seq) is one method that has revolutionised ecological, biogeographical and evolutionary studies (Hohenlohe *et al.*, 2010; Etter *et al.*, 2011; Andrews *et al.*, 2016). The ongoing development of whole-genome sequencing remains challenging compared to RAD-seq, which targets a reduced representation of the genomic regions flanking restriction sites (Baird *et al.*, 2008; Davey & Blaxter, 2010). RAD sequencing provides an efficient method for the discovery and genotyping of thousands of single nucleotide polymorphisms (SNPs) at sites scattered throughout the genome with no, or limited, prior genomic resources available for the organisms under study (Davey & Blaxter, 2010; Emerson *et al.*, 2010). Therefore, RAD-seq has been utilised for non-laborious and relatively cost-effective phylogenetic inference of model and non-model organisms (Cariou *et al.*, 2013; Suchan *et al.*, 2017; Wagner *et al.*, 2018). Raw RAD-seq reads can be mapped to a reference genome, if available, or processed *de novo* by clustering together reads based on a certain similarity threshold (CT). Despite recent advancements in *de novo* assembly pipelines (Willing *et al.*, 2011; Paris *et al.*, 2017; Díaz-Arce & Rodríguez-Ezpeleta, 2019), this approach still faces significant challenges, including sequencing errors, sequencing bias, repetitive region complexity and high computational requirements (Rubin *et al.*, 2012; Dida & Yi, 2021; Kunvar *et al.*, 2021). However, one of the critical considerations is that the final results of *de novo* assembly may strongly depend on the chosen filtering parameters, such as the selection of clustering threshold (Grzywacz *et al.*, 2021). On the other hand, studies have suggested that assembling raw RAD-seq reads to a reference genome can yield improved results compared to a *de novo* method, as it facilitates the determination of the genomic locations of loci and subsequently a higher number of SNPs calls (Manel *et al.*, 2016; Shafer *et al.*, 2017; Kunvar *et al.*, 2021). Reference sequences can be either the genome of the target species or a species closely related to the study group (Manel *et al.*, 2016). In some cases even a draft genome from a distant relative can be used with success (Shafer *et al.*, 2017; Kunvar *et al.*, 2021). However, an increase in evolutionary distance between ingroup taxa and reference genome can result in considerably lower phylogenetic signal and a failure in the reconstruction of relationships between deeper nodes (Tripp *et al.*, 2017; Grzywacz *et al.*, 2021).

Due to the unclear status of some of the proposed species groups within *Lispe*, this work aims to examine relationships within *Lispe* with the main objective to clarify the division of *Lispe* into species groups and analyse their phylogenetic relationships. To achieve this goal, and to investigate whether mapping short RAD-seq reads to a draft genome increases the phylogenetic signal, we applied RAD-seq using both *de novo* assembly and mapping to a reference genome under different analytical schemes. We utilise long reads sequencing to obtain a reference genome sequence of *L. tentaculata*.

2. Methods

2.1 Taxon sampling and DNA isolation

For phylogenetic analysis, we sampled 49 species of *Lispe* representing all recently proposed and/or revised species groups (Snyder, 1954; Hennig, 1960; Vikhrev, 2016, 2020, 2021, 2011a, 2011b, 2012a, 2012b, 2012c, 2014, 2015; Vikhrev *et al.*, 2016; Gao *et al.*, 2022) (Supp. Table S2). All adult specimens were identified by Nikita Vikhrev and AG using keys provided by Hennig (1955), Pont (2019) and Vikhrev (2020, 2021). Outgroups included four representatives of *Limnophora*. Voucher specimens, where available, have been deposited in the collection of the Department of Ecology and Biogeography, Faculty of Biological and Veterinary Sciences, Nicolaus Copernicus University in Toruń.

Prior to DNA extraction, ethanol-soaked samples were rinsed three times for 30 min in distilled water and dried on a thermoblock at 40□. Pinned specimens were directly used for DNA extraction. Total genomic DNA was isolated from entire specimens using a DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's protocol with the following modifications: (i) for each individual, 40 µL of proteinase K (>600 mAU/ml; Qiagen) was used; (ii) after initial incubation at 56°C, 4 µL of RNase A (100 mg/ml; Qiagen) was added to each sample. The extracted DNA was quantified with a Qubit 3.0 fluorometer using a dsDNA High Sensitivity Assay Kit (Life Technologies, Inc., Carlsbad, CA, USA) following the manufacturer's protocol. Samples with low DNA yield were additionally amplified with the REPLI-g Mini Kit (Qiagen) to increase DNA concentration. Extractions were electrophoresed in a 1% agarose gel, stained with GelRed (Biotium, Darmstadt, Germany) and photographed with a gel documentation system.

Specimens of *Lispe tentaculata* for nanopore sequencing were collected in Toruń, Poland (53°00'14.4"N 18°36'19.2"E) in June of 2021. Adults were placed in a freezer for a few minutes for immobilization. The material was subjected to DNA extraction, as described above. The two *L. tentaculata* isolates with the highest concentration and the longest DNA

fragments were selected based on results from the Qubit 3.0 assay and electrophoresis. The samples were subsequently purified with AMPure XP beads (Beckman Coulter, Carlsbad, CA, USA; 0.4 × ratio of beads to sample volume) to remove short DNA fragments and then re-suspended in TE buffer. The purified products were quantified with a Qubit 3.0 fluorometer using the dsDNA High Sensitivity Assay Kit following the manufacturer's protocol.

2.2 Library preparation for genome sequencing and data processing

Two libraries were prepared simultaneously, one for each *L. tentaculata* individual. We could not limit to one adult of *L. tentaculata* due to the low concentration of gDNA, insufficient for four sequencing runs on two flow cells. We used Ligation Sequencing Kit SQK-LSK110 (Oxford Nanopore Technologies, Oxford, United Kingdom) to prepare libraries according to the manufacturer's protocol with the following modifications: to minimise pipetting steps, the input DNA was prepared by transferring 1 µg of gDNA into a 0.2 ml PCR tube and adjusted with nuclease-free water to 47 µl. Times recommended for the initial 65°C binding incubation, incubation during a bead-based AMPure XP clean-up after DNA repair, end-prep, and adapter ligation steps, as well as incubation with a NEBNext Quick T4 DNA Ligase (New England BioLabs Inc., Ipswich, MA, USA) were doubled. For the beads washing step that follows the adapter ligation, the Long Fragment Buffer (LFB) was used. The incubation in the increased volume of 26 µl of Elution Buffer (EB) was performed at 37°C. After the completion of the library preparation protocol, half of the library (12 µl) was loaded onto a SpotON Flow Cell Rev D (R.9.4.1; FLO-MIN106D), while the other half and the second library were stored in 4°C according to the manufacturer's recommendations. After the first run, the flow cell was washed with the Flow Cell Wash Kit (EXP-WSK002) following the manufacturer's protocol and the second half of the library was immediately loaded for sequencing. A total of four sequencing runs were performed on two flow cells using a MinION Mk1C device (MIN-101C) (Oxford Nanopore Technologies, Oxford, United Kingdom).

The FAST5 ONT (Oxford Nanopore Technologies) reads were basecalled using *Guppy* v.5.0.7 with the super-accuracy mode (SUP). Further processing was preceded by the concatenation of all single fastq files into one fastq file. Adapter sequences were trimmed using *PoreChop* v.0.2.4 (https://github.com/rrwick/Porechop) (Wick *et al.*, 2017) with an option to discard reads with internal adapters. Next, reads were additionally filtered using

NanoFilt (https://github.com/wdecoster/nanofilt) and sequences shorter than 500 nucleotides or with Phred quality scores (Q) below 10 were removed. The quality check of overall raw and trimmed ONT reads was performed using *NanoPack* scripts (De Coster *et al.*, 2018).

In this study, five state-of-the-art *de novo* long-read only assemblers were utilised and, in an effort, to determine their efficiency and the completeness of assembled genomes. To assemble the *L. tentaculata* genome, we used: *Raven* v.1.5.1 (Vaser & Šikić, 2021), *SMARTdenovo* (Liu *et al.*, 2021), *wtdbg2* (Ruan & Li, 2020), *Canu* v.2.2 (Koren *et al.*, 2017) and *Flye* v.2.9-b1768 (Kolmogorov *et al.*, 2019). Default parameters were used in *Raven* assembler with two polishing rounds of *Racon* (Vaser *et al.*, 2017). In *SMARTdenovo* and *wtdbg2* assemblers, a minimum length of alignment was set to 1000 bp (-J and -L, respectively). For *wtdbg2*, *Canu* and *Flye* an approximate genome size was set to 700 Mb (Scott *et al.*, 2014). The accuracy and completeness of each *de novo* genome assembly was evaluated using *BUSCO* v.5.2.1 (Manni *et al.*, 2021). Analyses were performed with the odb10 Diptera lineage dataset from 56 genomes that was available in the NCBI GenBank in July 2022. Summary assembly statistics (number of contigs, total length, the longest contig, N50) and assembly quality (QV) were obtained using *Inspector* (Chen *et al.*, 2021). QV score was calculated based on the identified structural and small-scale errors scaled by the total base pairs of the assemblies. For subsequent analyses we selected the most complete assembly according to *BUSCO* assay and that with the highest QV.

2.3 RAD-seq library preparation and data processing

Genomic DNA for each species was individually barcoded and processed into a reduced complexity library based on the original RAD-seq protocol described by Ali *et al.* (2016) with the following modifications: (i) for each sample, two separate repetitions of 75 ng DNA each were digested using *Sbf*I-HF restriction enzyme (New England BioLabs) at 37°C for 2 hr to mitigate the risk of reaction failure; (ii) 5μL of P1 adapter-ligated fragments of each of the 106 samples (2 repetitions × 53 species) were pooled and then divided into three equal parts before the clean-up step; (iii) sonication was performed for 60 s using a Covaris M220 (Covaris, Inc. Woburn, MA, USA); (iv) Pippin Prep (Sage Science, Beverly, MA, USA) was used to select fragments between 300 and 500 bp with prior library cleaning with AMPure XP beads (1:1 ratio of beads to sample volume); (v) four independent PCRs (15 cycles) were carried out and subsequently pooled; and (vi) PCR products were purified twice with AMPure XP beads (1:1 ratio of beads to sample volume) to completely remove the remaining primers. A final library check was performed using a Qubit 3.0 fluorometer and 2100 Bioanalyzer with

the High Sensitivity DNA Analysis Kit (Agilent Technologies, Santa Clara, CA, USA). Commercial paired-end sequencing (Macrogen) of the multiplexed library was conducted using an Illumina HiSeq 2500.

Raw sequence read quality was analysed using *FastQC* v.0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/; accessed 27.03.2021). Illumina-specific adapters and low-quality bases were removed using *Trimmomatic* v.0.36 (Bolger *et al.*, 2014) with the following options: TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50. For downstream phylogenetic analyses, the raw RAD-seq R1 reads were processed with the *ipyrad* v.0.9.81 pipeline (Eaton, 2014). Reads were demultiplexed and assigned to species based on sequence barcodes (allowing one mismatch).

Further, two assembly pipelines were performed for 1) *de novo* assembly and 2) reference-guided assembly using the newly obtained reference genome sequence. For *de novo* assembly, various combinations of clustering thresholds (CT), i.e., the minimum percentage of sequences similarity below which two reads are considered to have come from different loci, were tested. Since this parameter is known to alter assembly results (Rubin *et al.*, 2012; Cariou *et al.*, 2013; Grzywacz *et al.*, 2021; Piwczyński *et al.*, 2021), we implemented a wide range of CT from 0.70 to 0.90 incremented by 0.01. Other parameters used for *ipyrad* analysis were as follows: *min_samples_locus* = 4, *max_SNPs_locus* = 0.6, *max_Indels_locus* = 8. For each alignment we performed phylogenetic tree reconstruction using maximum likelihood (ML) approach implemented in *RAxML* v.8.2.12 with 100 rapid bootstrap repetitions (Stamatakis, 2014). To select the best CT, we considered the highest average bootstrap support and the highest number of obtained SNPs. For reference-based approach, we performed assembly with a draft genome of *L. tentaculata* obtained in this study. The remaining parameters were kept unchanged, as previously mentioned.

2.4 Phylogenetic inference

Four alignments were analysed by maximum likelihood (ML) using *RAxML* v.8.2.6 (Stamatakis, 2014) under the concatenation approach: two *de novo* alignments with the highest average bootstrap support from preliminary study, one *de novo* alignment with the highest number of SNPs and one alignment obtained from mapping to reference genome. We applied nucleotide substitution model GTR + G. In the ML analysis, a search for the best scoring ML tree was performed with 100 replicates, and branch support for each node was assessed by standard 1000 nonparametric bootstrap replicates and summarised on the best ML tree.

We used the multispecies coalescent model as implemented in BPP v.4.0 software (Flouri *et al.*, 2018) to analyse RAD-seq data assembled with reference-based approach and under 0.74 CT. This approach allows investigation of the potential incomplete lineage sorting and can be used to account for sites in the genome that are evolutionarily linked which may lead to highly supported, yet incorrect species trees when analysed using concatenation-based approach. To infer a species tree using BPP we used A01 analysis (speciesdelimitation = 0 and speciestree = 1). We performed inference on two data sets: the first consisting of all 11 693 loci retrieved from reference-based assembly, and the second with 9 540 loci retrieved from assembly under a 0.74 similarity threshold. We conducted four independent MCMC runs with burn-in set to 10 000, a sample frequency of 5 and with 50 000 total samples. For each dataset we used the corresponding *RAxML* output as the starting species tree topology for the analysis. We specified inverse gamma priors for both population sizes (θ) and the divergence time of the root ($\tau 0$). We assigned the inverse gamma priors for θ with $\alpha = 3$ and $\beta = 0.02$, and for root age we set $\alpha = 3$ and $\beta = 0.234$. The divergence time between *Lispe* and *Limnophora* lineages was derived from Haseyama *et al.* (2015).

2.5 Data availability

Data obtained during this study have been submitted to NCBI (National Center for Biotechnology Information, Bethesda, MD, USA) and are available under the BioProject PRJNA1059801 accession number. Specifically, ONT long reads are available under SRR27397080 and RAD-seq reads under SRR27504576-SRR27504628 accession numbers in Sequence Read Archive (SRA). *Lispe tentaculata* Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JBBFKM000000000. The version described in this paper is version JBBFKM010000000.

3. Results

3.1 Nanopore sequencing and genome assemblies

Long-read sequences of *L. tentaculata* were obtained in two sequencing runs on MinION Mk1C. A total of 9.06×10^5 reads (6.01 Gb) were generated for the first run and 1.34×10^6 (5.3 Gb) for the second run. The N50 read lengths were 9.4 kb and 9.2 kb, respectively. We benchmarked five *de novo* tools for genome assembly from long ONT-only reads. The evaluation of genome completeness using *BUSCO* (Table 1; Fig. 2) revealed that *Flye* yielded the lowest percentage of missing genes (6.3%), with 90.1% of complete and 3.6% of fragmented genes. The lowest percentage of genome completeness values were observed in

Raven (49.9%) and *SMARTdenovo* (60.9%). *SMARTdenovo* generated the lowest assembly length (~282 Mb), followed by *Raven* (~335 Mb). *wtdbg2* resulted in an intermediate total length (~550 Mb), but the most contiguous assembly (N50 = 81.7 kb). *Flye* produced the longest assembly length of 989 Mb, yet the assembly was highly fragmented with the largest number of contigs (46 646) and the lowest N50 value (35.1 kb). Similarly, *Canu* produced a low contiguity assembly (N50 = 38.4 kb) with a total length of 748.8 Mb. The N50 value is commonly used as a parameter to reflect the contiguity of the assembly results. However, a larger N50 value is not always a useful parameter for assessing assembler performance, because longer contigs may be less accurate (Wang *et al.*, 2021). The Quality Value (QV), which estimates overall assembly quality, was the highest for *Flye* (30.97), followed by comparable values for *Canu* (25.43) and *SMARTdenovo* (24.89). *wtdbg2* showed the lowest QV (20.59). In general, *Raven* was the assembler with the lowest memory and computational requirements with the trade-off of poor statistical report in comparison to the other long-read assemblers. *Canu* and *Flye* performed better than the other three assemblers, not only by generating the highest assembly length but also having the highest QV and completeness. We selected the *Flye* assembler for the downstream analysis based on the *BUSCO* assessment (Fig. 2) and QV (Table 1).

3.2 RAD-seq assembly results

The highest number of parsimony informative sites (PIS) in alignments, a proxy of phylogenetic signal strength, was obtained under 0.74 CT (Supp. Fig. S2), while the highest mean bootstrap support (BS) was obtained under 0.85 CT (BS = 93.75%) and 0.75 CT (BS = 93.42%) (Supp. Fig. S1). Alignments obtained under these three CT were used for the subsequent phylogenetic analyses. The summary statistics for alignments obtained under 0.74 and 0.75 clustering thresholds were relatively similar, but the 0.74 CT obtained the highest mean bootstrap support (96%) (Table 2). The lowest number of retrieved loci (8 877), variable sites (113 339), PIS (28 234) and the lowest mean bootstrap value (84.2%) were recovered under 0.85 CT. Mapping reads to the *L. tentaculata* reference genome resulted in increased alignment length (1 786 362 bp), the number of retrieved loci (11 693) and variable sites (197 320), but simultaneously produced the lowest number of PIS (27 250) of all assembly methods (Table 2) and a mean bootstrap support of 87.7%.

The ML topologies produced varied between assembly methods particularly in terms of nodal support. The percentage of nodes with poor support (BS < 75%) varied, with 25% for the reference-based approach, 4% for *de novo* 0.74 CT, 15% for *de novo* 0.75 CT and 31% for *de*

novo 0.85 CT (Fig. 3; Supp. Fig. S3). Most other nodes resolved with high nodal support values (BS > 90%) or moderate support (75% < BS < 90%). In contrast, the topologies produced from each assembly method were generally congruent (Fig. 3; Supp. Fig. S3), except for the placement of *L. pumila* Wiedemann, 1824 and *L. pygmaea* and the position of the *leucospila*-group. These taxa resolved in two alternative topologies, one for both the *de novo* 0.74 (Fig. 3) and 0.75 CT (Supp. Fig. S3: topology A) assemblies, and the other for the reference-based (Fig. 3) and *de novo* 0.85 CT (Supp. Fig. S3: topology B) assembly analyses.
Our results are henceforth primarily described for the topologies derived from the reference-based and *de novo* 0.74 CT assemblies, as these are the datasets with the highest number of loci/PIS and highest mean bootstrap support, respectively (Fig. 3; Table 2). Subsequent descriptions of the bootstrap values for reference-based, *de novo* 0.74 CT, *de novo* 0.75 CT and *de novo* 0.85 CT assemblies are as follows BS_{rb}, $BS_{0.74}$, $BS_{0.75}$ and $BS_{0.85}$.

3.3 Concatenated maximum likelihood phylogenies

Our results, similarly to previous studies (Kutty *et al.*, 2010; Ge *et al.*, 2016; Grzywacz *et al.*, 2021; Gao *et al.*, 2022), confirm the monophyly of *Lispe* (Fig. 3; Supp. Fig. S3). A division into three highly supported (BS > 90%) clades is observed, both for reference-based and *de novo* assemblies. For the sake of transparency in the presentation of results and discussion, 'Clade A' includes the *palposa*-group, the *rigida*-group and the *caesia*-group; 'Clade B' is composed of *L. albimaculata* Stein, 1910 (not assigned to a group), the *nicobarensis*-group, the *nivalis*-group, the *scalaris*-group and the *tentaculata*-group; while 'Clade C' consists of the *longicollis*-group, the *desjardinsii*-group, the *uliginosa*-group and the *kowarzi*-group.
In the reference-based and *de novo* 0.85 CT assemblies, the *leucospila*-group is revealed to be the sister group to all other *Lispe* (BS_{rb} = 100%; $BS_{0.85}$ = 100%) (Fig. 3; Supp. Fig. S3: topology B), while in the *de novo* 0.74 and 0.75 CT assemblies the *leucospila*-group is a sister taxon of (*kowarzi*-group + (*uliginosa*-group + (*longicollis*-group + *desjardinsii*-group))) ($BS_{0.74}$ = 99%; $BS_{0.75}$ = 94%) (Fig. 3; Supp. Fig. S3: topology A). In *de novo* assembly under 0.74 CT (Fig. 3), a clade consisting of *palposa*-group, *rigida*-group and *caesia*-group ($BS_{0.74}$ = 100%) with *L. pumila* at the base ($BS_{0.74}$ = 73%) emerges as a sister group to the remaining *Lispe* ($BS_{0.74}$ = 100%).
Two species groups, the *pumila*-group and the *pygmaea*-group, each represented by a single representative, significantly differed in their position on the obtained phylogenetic trees. In the analysis of the reference-based assembly *L. pumila* is a sister taxon to *L. pygmaea* and this clade emerges a sister to the Clade B (Fig. 3). In *de novo* assemblies *L. pumila* and *L.*

pygmaea do not form a monophyletic clade, and the former is sister to Clade A, while the latter is sister to Clade B (Fig. 3).

The majority of the relationships within species groups are highly supported for the *leucospila*-, *longicollis*-, *nicobarensis*-, *nivalis*-, *rigida*-, *scalaris*-, *tentaculata*- and *uliginosa*-groups, and moderately or poorly supported for the *kowarzi*-, *caesia*- and *palposa*-groups (Fig. 3; Supp. Fig. S3). In Clade A, the *palposa*-group is monophyletic with *L. flavinervis* (Becker, 1904) being sister (BS_{rb} = 87%) to (*L. neimongola* Tian et Ma, 2000 + (*L. superciliosa* Loew, 1861 + *L. loewi* Ringdahl, 1922)) and these species form a sister group to *L. apicalis comitata* (Becker, 1904) + *L. apicalis apicalis* Mik, 1869 (BS_{rb} = 66%). Alternatively, in the *de novo* assemblies *L. apicalis comitata* + *L. apicalis apicalis* is sister to the rest of the *palposa*-group ($BS_{0.74}$ = 97%). The *rigida*-group either emerges as the sister group to the *palposa*-group (BS_{rb} = 100%) or is paraphyletic with regard to the *palposa*-group (Fig. 3 *de novo* CT: 0.74; Supp. Fig. S3). *Lispe cana* (Walker, 1849), traditionally representing the *cana*-group, is nested within the *caesia*-group, and emerges as a sister taxon to *L. flavicornis* Stein, 1909 (BS_{rb} = 38%) or to (*L. flavicornis* + (*L. polonaise* Vikhrev, 2021 + *L. caesia*)) ($BS_{0.74}$ = 93%). The *caesia*-group with *L. cana* is sister to the *palposa*- + *rigida*-groups (BS_{rb} = 99%; $BS_{0.74}$ = 100%).

In Clade B, *L. nana* Macquart, 1835, traditionally representing the *nana*-complex, is nested with *L. capensis* Zielke, 1971 (BS_{rb} = 100%) within the *tentaculata*-group, and *L. mirabilis* Stein, 1918 emerges as a sister taxon to *L. emdeni* Vikhrev, 2012 (BS_{rb} = 100%). The *nicobarensis*-group is sister to the *nivalis*-group (BS_{rb} = 100%; $BS_{0.74}$ = 99%). The *scalaris*-group is sister to (*L. albimaculata* + (*nigrimana* + *nivalis*groups)) with high branch support (BS_{rb} = 92%), and this clade is sister to the *tentaculata*-group (BS_{rb} = 90%) or to the (*nigrimana* + *nivalis*) groups with moderate support ($BS_{0.74}$ = 80%). *Lispe albimaculata* is sister taxon to the (*nicobarensis* + *nivalis*) groups (BS_{rb} = 67%) or to the (*scalaris* + (*nicobarensis* + *nivalis*)) groups ($BS_{0.74}$ = 99%).

In Clade C, a moderately supported dichotomy (BS_{rb} = 79%; $BS_{0.74}$ = 84%) splits the *longicollis*-group into subgroup I with (*L. longicollis* Meigen, 1826 + (*L. xenochaeta* Malloch, 1923 + *L. confusa* Vikhrev, 2021)) (BS_{rb} = 100%),and subgroup II with (*L. glabra* Wiedemann, 1824 + (*L. assimilis* Wiedemann, 1824 + *L. nuba* Wiedemann, 1830)) (BS_{rb} = 100%; $BS_{0.74}$ = 100%). *Lispe desjardinsii* is a sister taxon to the *longicollis*-group (BS_{rb} = 100%; $BS_{0.74}$ = 100%). The *uliginosa*-group is sister to the (*longicollis* + *desjardinsii*) groups , with moderate or high support (BS_{rb} = 63%; $BS_{0.74}$ = 92%). The poorly (BS_{rb} = 69%) or highly ($BS_{0.74}$ = 90) supported *kowarzi*-group is revealed as a sister to the traditionally

separated *geniseta*-complex represented by *L. geniseta* (BS_{rb} = 45%), and the former *dichaeta*-complex (*L. dichaeta* + *L. madagascariensis*) is sister (BS_{rb} = 69%) to the *kowarzi*-group + *L. geniseta*. In the *de novo* 0.74 CT topology, *L. geniseta* emerged as a sister taxon to the remaining representatives of *kowarzi*-group and the former *dichaeta*-complex ($BS_{0.74}$ = 99%).

3.4 Multispecies coalescence-based phylogenies

Assembly type did not influence the final BPP topologies, with both reference-based and *de novo* + 0.74 CT data sets producing congruent phylogenetic trees (Fig. 4). For the reference-based assembly, all species groups were monophyletic with maximum support (PP = 1). For the *de novo* + 0.74 CT assembly, similarly to the ML phylogenetic tree the *rigida*-group is paraphyletic with regard to the *palposa*-group (PP = 1).

The resultant phylogenetic trees from the BPP analysis were also highly congruent with the ML topologies obtained for reference-based and *de novo* + CT 0.74 assemblies, including the *leucospila*-group sister to remaining *Lispe*. *Lispe albimaculata* is sister to (*nivalis* + *nicobarensis*) groups (PP = 0.52 and 0.41 for reference and *de novo* assembly, respectively).

Alternatively, both reference-based and *de novo* + CT 0.74 BPP topologies disagree with their corresponding ML topologies in terms of the position of the clade *L. pumila* + *L. pygmaea*, which under BPP is sister to the remaining *Lispe* (PP = 1), with the exception of the *leucospila*-group.

4. Discussion

4.1 Nanopore sequencing performance

Many projects have recently been launched with the aim of providing high-quality genomes, e.g., Darwin Tree of Life, the Bird 10,000 Genomes (B10K) Project, the Vertebrate Genomes Project (VGP) and the BAT1K Genome Project. To date, the number of available dipteran genomes in public repository databases clearly indicates that model species (e.g., of *Drosophila* Fallén, 1823) or economically important species (e.g., of *Anopheles* Meigen, 1818) are of great interest. Among the Muscidae, genomes are available for ten species, that is *Eudasyphora cyanicolor* (Zetterstedt, 1845), *Haematobia irritans* (Linnaeus, 1758), *Hydrotaea cyrtoneurina* (Zetterstedt, 1845), *Hydrotaea diabolus* (Harris, 1780), *Musca domestica* Linnaeus, 1758, *Musca vetustissima* Walker, 1849, *Muscina levida* (Harris, 1780), *Phaonia tiefii* (Schnabl, 1888), *Polietes domitor* (Harris, 1780) and *Stomoxys calcitrans* (Linnaeus, 1758) (Scott *et al.*, 2014; Konganti *et al.*, 2018; Olafson *et al.*, 2021; Romine *et*

al., 2022; Falk & Grzywacz, 2024a, 2024b; Falk *et al.*, 2024), which represent a small percentage, considering that Muscidae are known from approximately 6 000 species. Nonetheless, this number is increasing especially due to the Darwin Tree of Life initiative.

Genome sequencing significantly advances phylogenetic research by providing extensive genetic data that enhances the resolution phylogenetic trees. This capability enables researchers to reconstruct phylogenetic relationships with greater comprehensiveness and reliability (McCormack *et al.*, 2013; Shakya *et al.*, 2020). In light of the progressive reduction of costs and computational requirements, genome sequencing is now achievable and affordable for individual laboratories, rather than only for international consortia (Brandies *et al.*, 2019). Despite only a few years of commercial use, nanopore sequencing has revolutionised genomic studies owing to facilitating *de novo* genome assembly by increasing read length and significantly reducing sequencing time (Leggett & Clark, 2017; Senol Cali *et al.*, 2018; *Nature*, 2023). Importantly, ONT sequencing is PCR-free, thereby avoiding biases during library preparation and mitigating the issues with assembling repetitive genome regions (Jansen *et al.*, 2017). However, as the availability of genome sequencing technology has increased, attention has also shifted to the pivotal step of genome assembly. This process may produce different results, depending on the use of various assemblers, each of which has its own algorithms and methodologies (Guiglielmoni *et al.*, 2021). To address crucial time and cost considerations, we used ONT reads to obtain the genome of *L. tentaculata*, evaluating five different assemblers, commonly used at the time of the study. Among the selected assemblers, *Flye* appeared to be the most effective, achieving the highest percentage of completeness (90.1%) in the *BUSCO* assessment and demonstrating the lowest small-scale assembly error rate per megabase of genome. Additionally, *Flye* exhibited the highest Quality Value (QV), indicating high genome reconstruction accuracy. Following *Flye*, *Canu* also showcased competitive performance, particularly in terms of completeness and assembly accuracy. This is in agreement with recent studies that compared different assemblers and in most of them *Flye* and *Canu* performed best, on both eukaryotic and prokaryotic genomes (Jung *et al.*, 2020; Latorre-Pérez *et al.*, 2020; Sun *et al.*, 2021; Cosma *et al.*, 2023). On the other hand, in this study *Raven* appears to perform relatively poorer compared to the other assemblers, showing lower completeness percentages and quality values, whereas in other genome comparison studies, it was noted as the best-performing assembler (Chen *et al.*, 2020). As previous studies have shown, there is no single assembler that stands out as the best, as various assemblers exhibit differences in terms of structural accuracy, completeness and contiguity of assembled genomes (Wick & Holt, 2019). This is due to the use of distinct

algorithms, optimisations for different data types and qualities and varied approaches to handling genomic complexity and error correction. Additionally, their performance can depend on specific settings, computational resources and ongoing software updates (Cosma *et al.*, 2023). Hence, the selection of the most appropriate assembler should depend on study-specific factors, including genome assembly goals such as the characteristics of the sequencing data and the complexity of the genome being studied. Furthermore, given the ongoing introduction of new assemblers and improvements to existing ones, it is advisable for users to keep updated with these advancements to ensure optimal performance that meets their specific needs.

Genomes of Muscidae, similarly to those of many other insects (Hotaling *et al.*, 2021), contain a large proportion of transposable elements. In case of muscid flies even more than 50% of the genome is present as repeated content (Romine *et al.*, 2022). The first genome of *Musca domestica* obtained with short reads approach allowed to assemble 691 Mb genome (GCA_000371365.1), while application of long reads sequencing allowed to overcome the issue of highly repetitive regions and assembly genomes ranging from 907 Mb (GCA_030504385.2) to 1.3 Gb (GCA_032878625.1) length. Among all the available genomes, only the genome of *Musca vetustissima* and the one obtained in this study for *Lispe tentaculata* were sequenced using nanopore sequencing, with the genome of the former species also incorporating short reads from Illumina. The remaining genomes were predominantly sequenced using PacBio technology, with some supplemented by short Illumina reads. With a genome size of 989 Mb assembled using *Flye* (Table 1), *L. tentaculata* falls in the mid-range of genome sizes among all available muscid genomes. It is larger than several genomes, like *H. cyrtoneurina* (575 Mb) and *M. vetustissima* (850 Mb), but smaller than others, such as those of *E. cyanicolor* and *P. tiefii*, which exceed 1.5 Gb.

4.2 Systematics of Lispe

The definition and species groups limits within *Lispe* have primarily relied on morphological data, lacking formal phylogenetic reconstruction. Gao *et al.* (2022) provided the first molecular phylogeny, revealing four major clades, but with low backbone support. In this study, we consistently observed the following three clades: Clade A (*palposa*-, *rigida*- and *caesia*-groups), Clade B (*nicobarensis*-, *nivalis*-, *scalaris*- and *tentaculata*-groups) and Clade C (*longicollis*-, *desjardinsii*-, *uliginosa*- and *kowarzi*-groups). Clade B and Clade C of this study are congruent in terms of species group composition with the second and third clades of Gao *et al.* (2022), but the limits of the remaining clades are incongruent between both studies. All incongruences between our generated topologies are related to relationships between these

three clades. None of our phylogenetic hypotheses are considered conclusive and future research is still needed to comprehensively resolve relationships in the backbone of the *Lispe* tree of life. Despite this, we can certainly review of the state of *Lispe* phylogenetics in light of our current results.

4.2.1 Relationships within Clade A

The *palposa*-group, originally proposed by Snyder (1954), is the most clearly defined group based on adult morphology and is closely related to the *rigida*-group, as concluded by Vikhrev (2015) and this study (BS = 100%). Neither the *leucospila*-group nor the *pygmaea*-group was found to be sister to the *palposa*-group as shown by Gao *et al.* (2022). The *caesia*-group, characterised by a widened ocellar triangle with convex margins, ventral spines on fore and mid femora and abdomen with a characteristic pattern, was one of the better-supported groups within *Lispe* according to Hennig (1960). Since then, species within the *caesia*-group have undergone re-examination, leading to a redefinition of this group by Gao *et al.* (2022). The present criterion for classifying species within the *caesia*-group is the presence of at least one of the character states indicated by Hennig. In line with this, *L. polonaise*, included for the first time in a molecular analysis, is nested within the *caesia*-group despite exhibiting only a slightly widened ocellar triangle with slightly convex margins (Vikhrev, 2021). *Lispe cana*, previously classified in the *cana*-group (Pont, 2019), is also nested within the *caesia*-group in our analysis, suggesting the inclusion of *L. cana* and its relatives within this group (Vikhrev, 2020). Our analyses revealed that the *caesia*-group is closely related to the (*uliginosa* + *rigida*) groups and it is not found to be sister to all other *Lispe* species as reported by Gao *et al.* (2022).

4.2.2 Relationships within Clade B

The systematic position of the *nana*-group and *L. mirabilis* within the *tentaculata*-group is congruent with previous molecular study (Gao *et al.*, 2022), thereby supporting the extended *tentaculata*-group sensu Gao *et al.* (2022). Additionally, we propose to include *L. capensis* in the *tentaculata*-group, as it is placed as sister to *L. nana* with very high support. This is in agreement with Vikhrev (2021), who stated that the intermediate character states of *L. capensis* support a relationship of *L. nana* with the *tentaculata*-group.

Our analyses show a sister-group relationship between the clade composed of *scalaris*-group and (*nivalis* + *nicobarensis*) groups, as well as the *tentaculata*-group, with variable nodal support (Fig. 3; Supp. Fig. S3). These results, in conjuction with Gao *et al*. (2022), support

Vikhrev's (2012a, 2014) conclusion that the *tentaculata* supergroup includes species from the *nivalis*-, *scalaris*- and extended *tentaculata*-groups (Vikhrev, 2014). Furthermore, we also propose to include the *nicobarensis*-group within this supergroup.

4.2.3 Relationships within Clade C

We confirm the split of the *longicollis*-group into subgroup I and II, as proposed by Hennig (1960), and further expanded on by Vikhrev (2014, 2020, 2021), with moderate support (BS_{rb} = 79%; $BS_{0.74}$ = 84%). Gao *et al*. (2022) suggested that *L. pennitarsis*, the only representative of the *desjardinsii*-group in their study, is nested within the *longicollis*-group. However, while *L. pennitarsis* appeared the intermediate position between the two subgroups of the *longicollis*-group (*assimilis*-subgroup and *longicollis*-subgroup), this position lacked significant support. Nevertheless, authors proposed merging the *desjardinsii*-subgroup into the *longicollis*-group. Despite morphological similarities between the *desjardinsii*- and *longicollis*-groups (Vikhrev, 2014), our study does not support the relationships reported by Gao *et al*. (2022). In all our analyses, *L. desjardinsii* is placed as a sister to the *longicollis*-group with full support (BS = 100%), not nested within it. We propose retaining the *desjardinsii*-group separately until more extensive taxon sampling is implemented to test the validity of the entire species group. Our study also shows that the *uliginosa*-group is sister to the *longicollis*-group + *desjardinsii*-group, which in turn is sister to the *kowarzi*-group sensu Gao *et al*. (2022). This is in conflict with Gao *et al*. (2022) that showed the *longicollis*-group (including the *desjardinsii*-group) sister to the *uliginosa*-group + *kowarzi*-complex. As for the former *geniseta* and *dichaeta* complexes, our results are congruent with those of Gao *et al.* (2022), which showed that these complexes clustered as sisters to the *kowarzi*-group. Therefore, we support the proposal of Gao *et al*. (2022) to extend the *kowarzi*-group to include all species assigned to the *dichaeta*, *geniseta* and *kowarzi* complexes.

4.2.4 Uncertain relationships of L. pumila, L. pygmaea *and* L. leucospila

The placement of *L. pumila* and *L. pygmaea* was influenced by the assembly method in our results. In the reference-based analysis, *L. pygmaea* emerged as sister to *L. pumila*, with moderate support (Fig. 3: reference-based), while the *de novo* + 0.74 CT analysis separated these species between the main three clades (Fig. 3: *de novo*, 0.74; Supp. Fig. S3). These two species were initially classified together in the *pumila*-group by Vikhrev (2012b), but later considered as monotypic complexes by the same author (Vikhrev, 2016), who proposed five complexes (*ambigua*, *dichaeta*, *geniseta*, *pumila* and *pygmaea*) that were regarded as the *L.*

pygmaea ecological group based on a shared ecology. *Lispe pumila* was excluded from the analysis in the previous study (Gao *et al.*, 2022), and we do not have a reference for its position on the *Lispe* tree. Thus, future studies with greater sampling are necessary to confirm the relationships between the *pumila*- and *pygmaea*-groups within *Lispe*.

The present study does not resolve the relationship between the *leucospila*-group and the other groups. The systematic position of the *leucospila*-group differed between our analyses. In in the reference-based approach it is sister to all other *Lispe* (Fig. 3: BS_{rb} = 100%), while in the *de novo* approach it is sister to the clade consisting of the *longicollis*-, *desjardinsii*-, *uliginosa*- and *kowarzi*-groups (Fig. 3: $BS_{0.74}$ = 99%; Supp. Fig. S3). Ge *et al.* (2016) also proposed that *L. leucospila* is at the base of the *Lispe* tree of life with high nodal support (BS = 98, PP = 1.0), however this study had incomplete taxon sampling, only including representatives of the *nivalis*-, *palposa*- and *tentaculata*-groups have been included (fig. 12 in Ge *et al.* 2016). Our results reject previous results by Gao *et al*. (2022) of a close relationship between the *leucospila*- and *palposa*-groups.

4.3 Ecology of Lispe

Species of the *leucospila*-group differ ecologically from other *Lispe* species. According to Vikhrev (2014), 'their typical habitats are grassy lawns being seasonally or artificially watered, or similar natural habitats, usually secondary sites with short or sparse grass and moderately wet soil', while other *Lispe* inhabit semi-aquatic environments with wet mud or sand and high organic content. While some species of *Lispe* can surely be distinguished by their ecology and are only encountered in selected habitats, the habitat preferences for many species are unknown or inconclusive. *Lispe sericipalpis* Stein, 1904 and *L. manicata* Wiedemann, 1830 were reported from limnic habitats (Vikhrev, 2011a, 2012c), while *L. candicans* and *L. caesia* were reported from habitats influenced by saltwater, and *L. orientalis* from habitats containing dirty and organically polluted water (Vikhrev, 2011a). In contrast, *L. pygmaea* and *L. uliginosa* were found in both limnic and saline habitats (Hennig, 1960). Therefore, using the ecology to infer phylogenetic relationships or ancestral state reconstruction should be approached with great caution. In this study we do not assess whether the habitat preferences observed within the *leucospila*-group, i.e., an association with grassy lawns and seasonally watered habitats, provide sufficient evidence to draw conclusions about the ancestral habitat preferences of *Lispe*. It is worth noting that many representatives of the closely related genus *Limnophora* are also associated with water bodies (Ivković & Pont, 2016). Thus, species of the *leucospila*-group may also exhibit a derived strategy, and an

association with stagnant or running water could potentially be an ancestral strategy within a larger clade comprising several muscid genera.

5. Conclusions

The results of this study are congruent with species-group concepts established using adult morphology, particularly those proposed by Vikhrev (2020, 2021). The findings of Gao *et al*. (2022), who provided the only previous phylogenetic hypothesis for the classification of *Lispe*, differed from our findings both in relationships between and within species groups. We propose expanding the *tentaculata* supergroup sensu Vikhrev (2014), which presently comprises the *tentaculata*-, *nivalis*- and *scalaris*-groups, to also include the *nicobarensis*-group. Our results corroborate the proposal of Gao *et al*. (2022) to expand the *kowarzi*-group to include the traditionally recognised *dichaeta*-complex and *geniseta*-complex. Given that our results provide strong support for *L. desjardinsii* as the sister taxon of the *longicollis*-group, we retain the validity of the *desjardinsii*-group, thus confirming the presence of 14 distinct species groups in the genus *Lispe*. Our results yielded two alternate, but highly supported phylogenetic tree topologies resolving most of the relationships between and within species groups and between species within those groups. Future studies focusing on the genus *Lispe* should prioritise improving taxon sampling, including both species from recognised species groups and many of those that have not yet been assigned to any group (Supp. Table S1).

Declaration of Competing Interest

The authors declare that they have no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Acknowledgements

We would like to express our appreciation to Nikita E. Vikhrev and Konstantin Tomkovich (Moscow, Russia), Adrian C. Pont (Oxford, UK), Alessandra Rung, Martin Hauser and Stephen D. Gaimari (Sacramento, CA, USA), Mike E. Irwin (Urbana-Champaign, IL, USA), Thai Hong Pham (Hanoi, Vietnam), Oleg Kosterin and Natalya Priydak (Novosibirsk, Russia) and Doreen Werner (Müncheberg, Germany) for help in obtaining material. This research was supported by the National Science Centre of Poland (grant no. 2019/33/B/NZ8/02316 to AG). We gratefully acknowledge Poland's high-performance computing infrastructure PLGrid

(HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2022/015448 to AG.

References

Ali, O.A., O'Rourke, S.M., Amish, S.J., Meek, M.H., Luikart, G., Jeffres, C., & Miller, M.R. (2016) RAD capture (Rapture): Flexible and Efficient Sequence-Based Genotyping. Genetics, 202, 389–400.

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., & Hohenlohe, P.A. (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet, 17, 81–92.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., & Johnson, E.A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE, 3, 1–7.

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

Brandies, P., Peel, E., Hogg, C.J., & Belov, K. (2019) The value of reference genomes in the conservation of threatened species. Genes, 10, 846.

Butterworth, N.J. & Wallman, J.F. (2022) Flies getting filthy: The precopulatory mating behaviours of three mud-dwelling species of Australian Lispe (Diptera: Muscidae). Ethology, 128, 369–377.

Cariou, M., Duret, L., & Charlat, S. (2013) Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. Ecology and Evolution, 3, 846–852.

Chen, Y., Zhang, Y., Wang, A.Y., Gao, M., & Chong, Z. (2021) Accurate long-read de novo assembly evaluation with Inspector. Genome Biology, 22, 1–21.

Chen, Z., Erickson, D.L., & Meng, J. (2020) Benchmarking long-read assemblers for genomic analyses of bacterial pathogens using oxford nanopore sequencing. International Journal of Molecular Sciences, 21, 1–27.

Cosma, B.M., Shirali Hossein Zade, R., Jordan, E.N., Van Lent, P., Peng, C., Pillay, S., & Abeel, T. (2023) Evaluating long-read de novo assembly tools for eukaryotic genomes: insights and considerations. GigaScience, 12, 1–12.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., & Van Broeckhoven, C. (2018) NanoPack: Visualizing and processing long-read sequencing data. Bioinformatics, 34, 2666–2669.

Curran, C.H. (1937) African Muscidae - IV (Diptera). American Museum Novitates, 1–14.

Davey, J.L. & Blaxter, M.W. (2010) RADseq: Next-generation population genetics. Briefings in Functional Genomics, 9, 416–423.

Díaz-Arce, N. & Rodríguez-Ezpeleta, N. (2019) Selecting RAD-Seq data analysis parameters for population genetics: The more the better? Frontiers in Genetics, 10, 1–10.

Dida, F. & Yi, G. (2021) Empirical evaluation of methods for de novo genome assembly. PeerJ Computer Science, 7, 1–31.

Eaton, D.A.R. (2014) PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics, 30, 1844–1849.

van Emden, F.I. (1941) Key to the Muscidae of the Ethiopian Region: Scatophaginae, Anthomyiinae, Lispinae, Fanniinae. Bulletin of Entomological Research, 32, 251–275.

van Emden, F.I. (1965) Diptera 7, Muscidae, Part I. The fauna of India and the adjacent countries. Government of India, Delhi, 647.

Emerson, K.J., Merz, C.R., Catchen, J.M., Hohenlohe, P.A., Cresko, W.A., Bradshaw, W.E., & Holzapfel, C.M. (2010) Resolving postglacial phylogeography using high-throughput sequencing. Proceedings of the National Academy of Sciences of the United States of America, 107, 16196–16200.

Etter, P.D., Bassham, S., Hohenlohe, P.A., Johnson, E.A., & Cresko, W.A. (2011) SNP Discovery and Genotyping for Evolutionary Genetics Using RAD Sequencing. Methods in Molecular Biology, 772, 157–178.

Falk, S. & Grzywacz, A. (2024a) The genome sequence of a muscid fly, Hydrotaea cyrtoneurina (Zetterstedt, 1845). Wellcome Open Research, 9, 60.

Falk, S. & Grzywacz, A. (2024b) The genome sequence of a muscid fly, Hydrotaea diabolus (Harris, [1780]). Wellcome Open Research, 9, 176.

Falk, S., Sivell, D., Webb, J., & Grzywacz, A. (2024) The genome sequence of a muscid fly, Polietes domitor (Harris, 1780). Wellcome Open Research, 9, 58.

Fan, Z.-D. (2008) Fauna Sinica Insecta. Diptera Muscidae (I), vol. 49. pp. 1–1180. Science Press, Beijing.

Flouri, T., Jiao, X., Rannala, B., & Yang, Z. (2018) Species tree inference with BPP using genomic sequences and the multispecies coalescent. Molecular Biology and Evolution, 35, 2585–2593.

Frantsevich, L. & Gorb, S. (2006) Courtship dances in the flies of the genus Lispe (Diptera: Muscidae): From the fly's viewpoint. Archives of Insect Biochemistry and Physiology, 62, 26–42.

Gao, Y., Ge, Y., Yan, L., Vikhrev, N.E., Wang, Q., Butterworth, N.J., & Zhang, D. (2022) Phylogenetic Analyses Support the Monophyly of the Genus Lispe Latreille (Diptera: Muscidae) with Insights into Intrageneric Relationships. Insects, 13, 1015.

Ge, Y., Gao, Y., Yan, L., Liu, X., & Zhang, D. (2016) Review of the Lispe tentaculata-group (Diptera: Muscidae) in China, with one new synonym. Zoosystema, 38, 339–352.

Grzywacz, A., Trzeciak, P., Wiegmann, B.M., Cassel, B.K., Pape, T., Walczak, K., Bystrowski, C., Nelson, L., & Piwczyński, M. (2021) Towards a new classification of Muscidae (Diptera): a comparison of hypotheses based on multiple molecular phylogenetic approaches. Systematic Entomology, 46, 508–525.

Guiglielmoni, N., Houtain, A., Derzelle, A., Van Doninck, K., & Flot, J.F. (2021) Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. BMC Bioinformatics, 22, 1–23.

Haseyama, K.L.F., Wiegmann, B.M., Almeida, E.A.B., & de Carvalho, C.J.B. (2015) Say goodbye to tribes in the new house fly classification: A new molecular phylogenetic analysis and an updated biogeographical narrative for the Muscidae (Diptera). Molecular Phylogenetics and Evolution, 89, 1–12.

Hennig, W. (1955) 63 b. Muscidae. In: Lindner E., editor. Die Fliegen der Paläarktischen Region. Stuttgart: Schweizerbart'sche Verlagsbuchhandlung. p. 625–672.

Hennig, W. (1960a) 63 b. Muscidae (Lieferung 225). In: Lindner E., editor. Die Fliegen der Palaearktischen Region. Stuttgart: Schweizerbart'sche Verlagsbuchhandlung. p. 399–432.

Hennig, W. (1960b) Muscidae (Lieferung 209). In: Lindner E., editor. Die Fliegen der Palaearktischen Region. Stuttgart: Schweizerbart'sche Verlagsbuchhandlung. p. 399–432.

Hennig, W. (1965) Vorarbeiten zu einem phylogenetischen System der Muscidae (Diptera: Cyclorrhapha). Stuttgarter Beiträge zur Naturkunde, 141, 1–100.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A., & Cresko, W.A. (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genetics, 6, e1000862.

Hotaling, S., Sproul, J.S., Heckenhauer, J., Powell, A., Larracuente, A.M., Pauls, S.U., Kelley, J.L., & Frandsen, P.B. (2021) Long Reads Are Revolutionizing 20 Years of Insect Genome Sequencing. Genome Biology and Evolution, 13,.

Ivković, M. & Pont, A.C. (2016) Long-time emergence patterns of Limnophora species (Diptera, Muscidae) in specific karst habitats: tufa barriers. Limnologica, 61, 29–35.

Jansen, H.J., Liem, M., Jong-Raadsen, S.A., Dufour, S., Weltzien, F.A., Swinkels, W., Koelewijn, A., Palstra, A.P., Pelster, B., Spaink, H.P., Thillart, G.E.V. Den, Dirks, R.P., & Henkel, C. V. (2017) Rapid de novo assembly of the European eel genome from

nanopore sequencing reads. Scientific Reports, 7, 1–13.

Jung, H., Jeon, M.S., Hodgett, M., Waterhouse, P., & Eyun, S. Il. (2020) Comparative Evaluation of Genome Assemblers from Long-Read Sequencing for Plants and Crops. Journal of Agricultural and Food Chemistry, 68, 7670–7677.

Karl, O. (1928) Zweiflügler oder Diptera. II: Muscidae. Die Tierwelt Deutschlands und der angrenzenden Meeresteile nach ihren Merkmalen und nach ihrer Lebensweise. p. 1–232.

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P.A. (2019) Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology, 37, 540–546.

Konganti, K., Guerrero, F.D., Schilkey, F., Ngam, P., Jacobi, J.L., Umale, P.E., Perez de Leon, A.A., & Threadgill, D.W. (2018) A whole genome assembly of the Horn Fly, Haematobia irritans, and prediction of genes with roles in metabolism and sex determination. G3 Genes|Genomes|Genetics, 8, 1675–1686.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., & Phillippy, A.M. (2017) Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Research, 27, 722–736.

Kunvar, S., Czarnomska, S., Pertoldi, C., & Tokarska, M. (2021) In Search of Species-Specific SNPs in a Non-Model Animal (European bison (Bison bonasus))—Comparison of De Novo and Reference-Based Integrated Pipeline of STACKS Using Genotyping-by-Sequencing (GBS) Data. Animals, 11, 1–13.

Kutty, S.N., Pape, T., Wiegmann, B.M., & Meier, R. (2010) Molecular phylogeny of the Calyptratae (Diptera: Cyclorrhapha) with an emphasis on the superfamily Oestroidea and the position of Mystacinobiidae and McAlpine's fly. Systematic Entomology, 35, 614–635.

Latorre-Pérez, A., Villalba-Bermell, P., Pascual, J., & Vilanova, C. (2020) Assembly methods for nanopore-based metagenomic sequencing: a comparative study. Scientific Reports, 10, 1–14.

Leggett, R.M. & Clark, M.D. (2017) A world of opportunities with nanopore sequencing. Journal of Experimental Botany, 68, 5419–5429.

Liu, H., Wu, S., Li, A., & Ruan, J. (2021) SMARTdenovo: a de novo assembler using long noisy reads. Gigabyte, 2021, 1–9.

Malloch, J.R. (1923) Notes on Australian Diptera with descriptions. [No. i.]. Proceedings of the Linnean Society of New South Wales, 48, 601–622.

Manel, S., Perrier, C., Pratlong, M., Abi-Rached, L., Paganini, J., Pontarotti, P., & Aurelle, D. (2016) Genomic resources and their influence on the detection of the signal of positive

selection in genome scans. Molecular Ecology, 25, 170–184.

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., & Zdobnov, E.M. (2021) BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution, 38, 4647–4654.

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., & Brumfield, R.T. (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution, 66, 526–538.

Metzker, M.L. (2010) Sequencing technologies - the next generation. Nature reviews. Genetics, 11, 31–46.

Nature. (2023) Method of the Year 2022: long-read sequencing. Nature Methods, 20,.

Olafson, P.U., Aksoy, S., Attardo, G.M., Buckmeier, G., Chen, X., Coates, C.J., Davis, M., Dykema, J., Emrich, S.J., Friedrich, M., Holmes, C.J., Ioannidis, P., Jansen, E.N., Jennings, E.C., Lawson, D., Martinson, E.O., Maslen, G.L., Meisel, R.P., Murphy, T.D., Nayduch, D., Nelson, D.R., Oyen, K.J., Raszick, T.J., Ribeiro, J.M.C., Robertson, H.M., Rosendale, A.J., Sackton, T.B., Saelao, P., Swiger, S.L., Sze, S.-H., Tarone, A.M., Taylor, D.B., Warren, W.C., Waterhouse, R.M., Weirauch, M.T., Werren, J.H., Wilson, R.K., Zdobnov, E.M., & Benoit, J.B. (2021) The genome of the stable fly, *Stomoxys calcitrans*, reveals potential mechanisms underlying reproduction, host interactions, and novel targets for pest control. BMC Biology, 19, 41.

Paris, J.R., Stevens, J.R., & Catchen, J.M. (2017) Lost in parameter space: a road map for stacks. Methods in Ecology and Evolution, 8, 1360–1373.

Paterson, H.E. (1953) New Lispe species (Dipt., Muscidae) from Southern Africa. Journal of the Entomological Society of Southern Africa, 16, 168–178.

Piwczyński, M., Trzeciak, P., Popa, M.O., Pabijan, M., Corral, J.M., Spalik, K., & Grzywacz, A. (2021) Using RAD seq for reconstructing phylogenies of highly diverged taxa: A test using the tribe Scandiceae (Apiaceae). Journal of Systematics and Evolution, 59, 58–72.

Pont, A.C. (1986) Family Muscidae. In: Soós A., Papp L., editors. Catalogue of Palaearctic Diptera. Scatophagidae - Hypodermatidae. Amsterdam: Elsevier. p. 1–345.

Pont, A.C. (2019) Studies on the Australian Muscidae (Diptera). VIII. The genus Lispe Latreille, 1797. pp. 1–232. .

Romine, M.G., Knutie, S.A., Crow, C.M., Vaziri, G.J., Chaves, J.A., Koop, J.A.H., & Lamichhaney, S. (2022) The genome sequence of the avian vampire fly (Philornis downsi), an invasive nest parasite of Darwin’s finches in Galápagos. G3

Genes|Genomes|Genetics, 12,.

Ruan, J. & Li, H. (2020) Fast and accurate long-read assembly with wtdbg2. Nature Methods, 17, 155–158.

Rubin, B.E.R., Ree, R.H., & Moreau, C.S. (2012) Inferring phylogenies from RAD sequence data. PLoS ONE, 7, 1–12.

Scott, J.G., Warren, W.C., Beukeboom, L.W., Bopp, D., Clark, A.G., Giers, S.D., Hediger, M., Jones, A.K., Kasai, S., Leichter, C.A., Li, M., Meisel, R.P., Minx, P., Murphy, T.D., Nelson, D.R., Reid, W.R., Rinkevich, F.D., Robertson, H.M., Sackton, T.B., Sattelle, D.B., Thibaud-Nissen, F., Tomlinson, C., van de Zande, L., Walden, K.K., Wilson, R.K., & Liu, N. (2014) Genome of the house fly, Musca domestica L., a global vector of diseases with adaptations to a septic environment. Genome Biology, 15, 466.

Séguy, E. (1937) Diptera, family Muscidae. In: P. Wytsmann (ed.). Genera Insectorum. p. 1-604 Bruxelles, Desmet-Verteneuil.

Senol Cali, D., Kim, J.S., Ghose, S., Alkan, C., & Mutlu, O. (2018) Nanopore sequencing technology and tools for genome assembly: Computational analysis of the current state, bottlenecks and future directions. Briefings in Bioinformatics, 20, 1542–1559.

Shafer, A.B.A., Peart, C.R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C.W., & Wolf, J.B.W. (2017) Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. Methods in Ecology and Evolution, 8, 907–917.

Shakya, M., Ahmed, S.A., Davenport, K.W., Flynn, M.C., Lo, C.C., & Chain, P.S.G. (2020) Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. Scientific Reports, 10, 1–15.

Shinonaga, S. & Kano, R. (1983) Two new species and a newly recorded subspecies of the genus Lispe Latreille from Japan with a key to Japanese species (Diptera, Muscidae). Medical Entomology and Zoology, 34, 83–88.

Skidmore, P. (1985) The biology of the Muscidae of the World. Series Entomologica, 29, 1–550.

Snyder, F.M. (1949) New genera and species of Lispinae (Diptera, Muscidae). American Museum novitates, 1403, 1–9.

Snyder, F.M. (1954) A review of Nearctic Lispe Latreille (Diptera, Muscidae). The American Museum of Natural History, 1–40.

Stamatakis, A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30, 1312–1313.

Steidle, J.L.M., Dettner, K., Hübner, G., Köpf, A., & Reinhard, J. (1995) The predaceous fly Lispe candicans (Diptera: Muscidae) and its chemically protected prey, the rove beetle Bledius furcatus (Coleoptera: Staphylinidae). Entomologia generalis, 20, 11–19.

Suchan, T., Espíndola, A., Rutschmann, S., Emerson, B.C., Gori, K., Dessimoz, C., Arrigo, N., Ronikier, M., & Alvarez, N. (2017) Assessing the potential of RAD-sequencing to resolve phylogenetic relationships within species radiations: The fly genus Chiastocheta (Diptera: Anthomyiidae) as a case study. Molecular Phylogenetics and Evolution, 114, 189–198.

Sun, J., Li, R., Chen, C., Sigwart, J.D., & Kocot, K.M. (2021) Benchmarking Oxford Nanopore read assemblers for high-quality molluscan genomes. Philosophical Transactions of the Royal Society B: Biological Sciences, 376,.

Tripp, E.A., Tsai, Y.H.E., Zhuang, Y., & Dexter, K.G. (2017) RADseq dataset with 90% missing data fully resolves recent radiation of Petalidium (Acanthaceae) in the ultra-arid deserts of Namibia. Ecology and Evolution, 7, 7920–7936.

Vaser, R. & Šikić, M. (2021) Time- and memory-efficient genome assembly with Raven. Nature Computational Science, 1, 332–336.

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research, 27, 737–746.

Vikhrev, N. (2011a) Review of the Palaearctic members of the Lispe tentaculata species-group (Diptera, Muscidae): Revised key, synonymy and notes on ecology. ZooKeys, 84, 59–70.

Vikhrev, N.E. (2011b) Taxonomic notes on the Lispe leucospila species-group (Diptera: Muscidae). Russian Entomological Journal, 20, 215–218.

Vikhrev, N.E. (2012a) Four new species of Lispe Latreille, 1796 (Diptera: Muscidae) with taxonomic notes on related species. Russian Entomological Journal, 21, 423–433.

Vikhrev, N.E. (2012b) Revision of the Lispe longicollis-group (Diptera, Muscidae). ZooKeys, 235, 23–39.

Vikhrev, N.E. (2012c) Notes on taxonomy of Lispe Latreille (Diptera: Muscidae). Russian Entomological Journal, 21, 107–112.

Vikhrev, N.E. (2014) Taxonomic notes on Lispe (Diptera, Muscidae), Parts 1-9. Amurian zoological journal, VI, 147–170.

Vikhrev, N.E. (2015) Taxonomic notes on Lispe (Diptera, Muscidae), Parts 10-12. Amurian zoological journal, 7, 228–247.

Vikhrev, N.E. (2016) Taxonomic notes on Lispe (Diptera, Muscidae), Part 13. Amurian

zoological journal, VIII, 171–185.

Vikhrev, N.E. (2020) Lispe (Diptera, Muscidae) of the Palaearctic region. pp. 1–158–188. .

Vikhrev, N.E. (2021) Lispe (Diptera, Muscidae) of Africa. pp. 1–369–400. .

Vikhrev, N.E., Ge, Y.Q., & Zhang, D. (2016) On taxonomy of the Lispe caesia-group (Diptera: Muscidae). Russian Entomological Journal, 25, 407–410.

Wagner, N.D., Gramlich, S., & Hörandl, E. (2018) RAD sequencing resolved phylogenetic relationships in European shrub willows (Salix K. subg. Chamaetia and subg. Vetrix) and revealed multiple evolution of dwarf shrubs. Ecology and Evolution, 8, 8243–8255.

Wang, J., Chen, K., Ren, Q., Zhang, Y., Liu, J., Wang, G., Liu, A., Li, Y., Liu, G., Luo, J., Miao, W., Xiong, J., Yin, H., & Guan, G. (2021) Systematic Comparison of the Performances of De Novo Genome Assemblers for Oxford Nanopore Technology Reads From Piroplasm. Frontiers in Cellular and Infection Microbiology, 11,.

Werner, D., Pont, A.C., & Kampen, H. (2014) Lispe tentaculata (De Geer) and Lispe pygmaea Fallén (Muscidae) as natural predators of mosquitoes (Culicidae), with a review of mosquito predation by Lispe Latreille. Studia Dipterologica, 21, 11–22.

Wick, R.R. & Holt, K.E. (2019) Benchmarking of long-read assemblers for prokaryote whole genome sequencing. F1000Research, 8, 1–22.

Wick, R.R., Judd, L.M., Gorrie, C.L., & Holt, K.E. (2017) Completing bacterial genome assemblies with multiplex MinION sequencing. Microbial Genomics, 3, e000132.

Williams, F.X. (1938) Biological studies in Hawaiian water-loving insects. Part III. Diptera or flies. A, Ephydridae and Anthomyiidae. Proceedings of the Hawaiian Entomological Society, 10, 85–119.

Willing, E.M., Hoffmann, M., Klein, J.D., Weigel, D., & Dreyer, C. (2011) Paired-end RAD-seq for de novo assembly and marker design without available reference. Bioinformatics, 27, 2187–2193.

Zielke, E. (2018) Revalidation of an erroneously synonymized Helina species and reminder of a forgotten Lispe species from the Afrotropical Region. Beiträge zur Entomologie = Contributions to Entomology, 68, 371–372.

Table 1. Evaluation summary of genome completeness (*BUSCO*) and genome assembly quality (Inspector) for five assemblers: *Raven*, *SMARTdenovo*, *wtdbg2*, *Canu* and *Flye*. *BUSCO* assessment used the dipteran dataset (3285 genes).

Parameter	*Raven*	*SMARTdenovo*	*wtdbg2*	*Canu*	*Flye*
***BUSCO*, n = 3285**					
Complete [single, duplicated]	49.9% [49.5%, 0.4%]	60.9% [60.1%, 0.8%]	77.3% [77.1%, 0.2%]	82.5% [60.8%, 21.7%]	90.1% [67.9%, 22.2%]
Fragmented	6.1%	5.7%	7.5%	3.8%	3.6%
Missing	44.0%	33.4%	15.2%	13.7%	6.3%
Inspector					
Number of contigs	6049	7870	14130	23668	46646
Number of contigs > 1000 bp	6049	7870	14127	23668	45777
Number of contigs > 10000 bp	6042	7595	9775	21765	28848
Total length	334979448 (335 Mb)	281692310 (281,7 Mb)	549512882 (549,5 Mb)	748791779 (748,8 Mb)	988956643 (989 Mb)
Longest contig	447123	795484	1985679	572497	384705
N50	68227	42005	81745	38417	35162
Mapping rate	85.06%	82.59%	94.02%	91.56%	93.77%
Split-read rate	44.76%	41.45%	41.43%	36.49%	38.95%
Depth	31.72	37.08	20.17	14.64	12.33
Small-scale assembly error/Mbp	3686.95	2355.19	7463.49	2362.22	719.94
Total small-scale assembly error	1235053	663439	4101272	1768812	645359
Quality Value (QV)	23.00	24.89	20.59	25.43	30.97

Table 2. Summary statistics of analysed data and summary of bootstrap support values for phylogenies inferred from restriction site associated DNA sequencing (RAD-seq) alignments with maximum likelihood approach. RAD-seq data were processed with *de novo* approach under 0.74, 0.75 and 0.85 clustering thresholds and with the reference-based approach with the genome sequence of *Lispe tentaculata*. Abbreviations: PIS, parsimony informative sites; CV, coefficient of variation.

							Bootstrap support		
Analysed data		Aligment (bp)	Loci	Missing data (%)	Variable sites	PIS	Mean	Median	CV
Reference-based	*Flye*	1 786 362	11 693	90.44	197 320	27 250	87.7%	100	0.20
de novo	Threshold								
	0.74	783 882	9 540	89.00	155 298	38 929	96%	100	0.08
	0.75	788 761	9 640	89.02	154 955	38 830	92.6%	100	0.14
	0.85	731 488	8 877	88.73	113 339	28 234	84.2%	100	0.27

Figure captions

Fig. 1 Representative taxa of *Lispe*. (A) *Lispe assimilis* Wiedemann, 1824; (B) *Lispe caesia* Meigen, 1826; (C) *Lispe loewi* Ringdahl, 1922; (D) *Lispe nana* Macquart, 1835; (E) *Lispe polonaise* Vikhrev, 2021; (F) *Lispe pygmaea* (Fallén, 1825); (G) *Lispe sydneyensis* Schiner, 1868; (H) *Lispe tentaculata* (De Geer, 1776). Scale bar 3 mm.

Fig. 2 *BUSCO* analysis for the completeness of the *Lispe tentaculata* genome assembly using *Canu*, *Flye*, *Raven*, *SMARTdenovo* and *wtbg2* against Diptera reference dataset. The y-axis indicates five assemblers used in this study, and the x-axis shows the percentage of complete and single-copy, complete and duplicated, fragmented and missing genes in assembled contigs.

Fig. 3 Comparison of *RAxML* tree topologies inferred from reference-based approach and *de novo* assembly under 0.74 clustering threshold. Node support values are shown for 1000 nonparametric bootstrap replicates (BS). Species groups of *Lispe* are marked with different colours as provided and outgroup marked with a black colour. Clades consistently observed in this study are indicated by specific colours. Clade A consists of the *palposa*-, *rigida*-, and *caesia* groups, Clade B includes the *nicobarensis*-, *nivalis*-, *scalaris*- and *tentaculata* groups and Clade C includes the *longicollis*-, *desjardinsii*-, *uliginosa*- and *kowarzi* groups. In this study, the placement of *L. pumila*, *L. pygmaea*, and *L. leucospila* was influenced by assembly method and therefore they were not assigned to any clade. Nodes with BS = 100 are marked with an asterisk (*). Abbreviation: CT, clustering threshold.

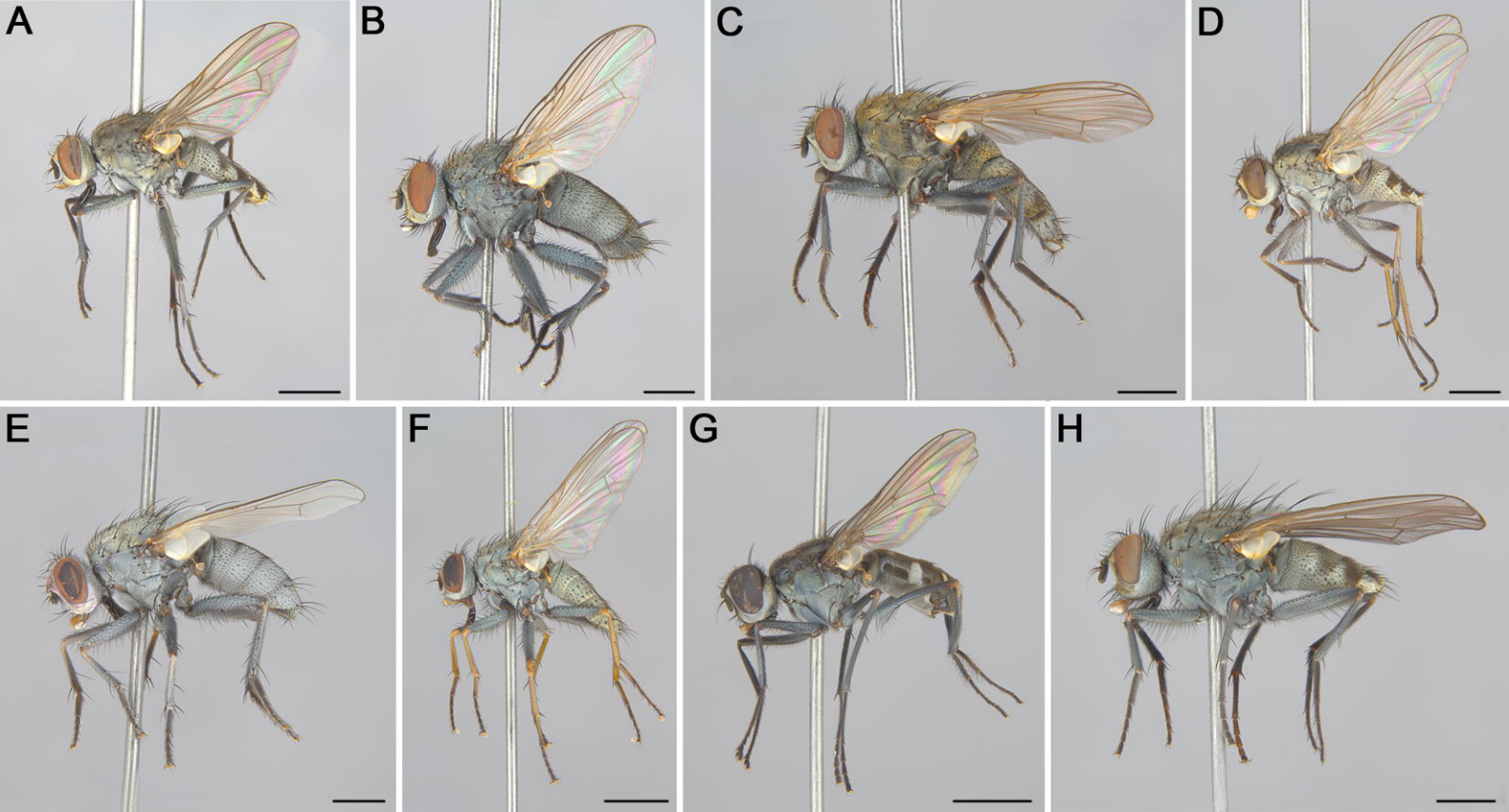
Fig. 4. Comparison of BPP trees topologies inferred from reference-based approach and *de novo* assembly under 0.74 clustering threshold clustering threshold. Species groups are marked with different colours and collapsed for monophyletic groups which received maximum node support (PP = 1). Outgroup marked with black colour include representatives of *Limnophora*. Abbreviation: CT, clustering threshold.

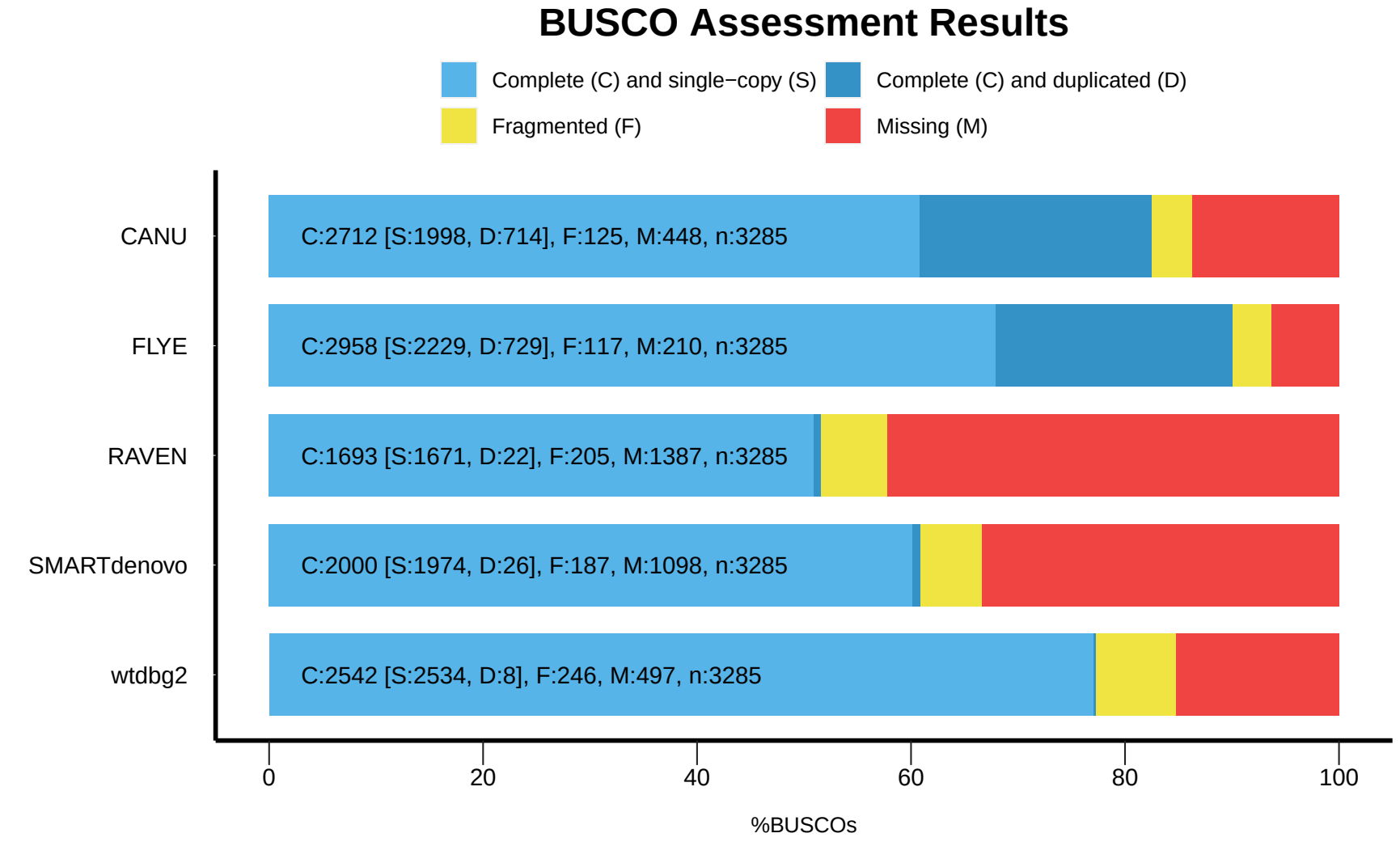
Fig. S1 Distribution of bootstrap values for preliminary maximum likelihood (ML) analysis followed with 100 rapid bootstrap repetitions for datasets obtained with *de novo* assembly under different clustering thresholds (CT).

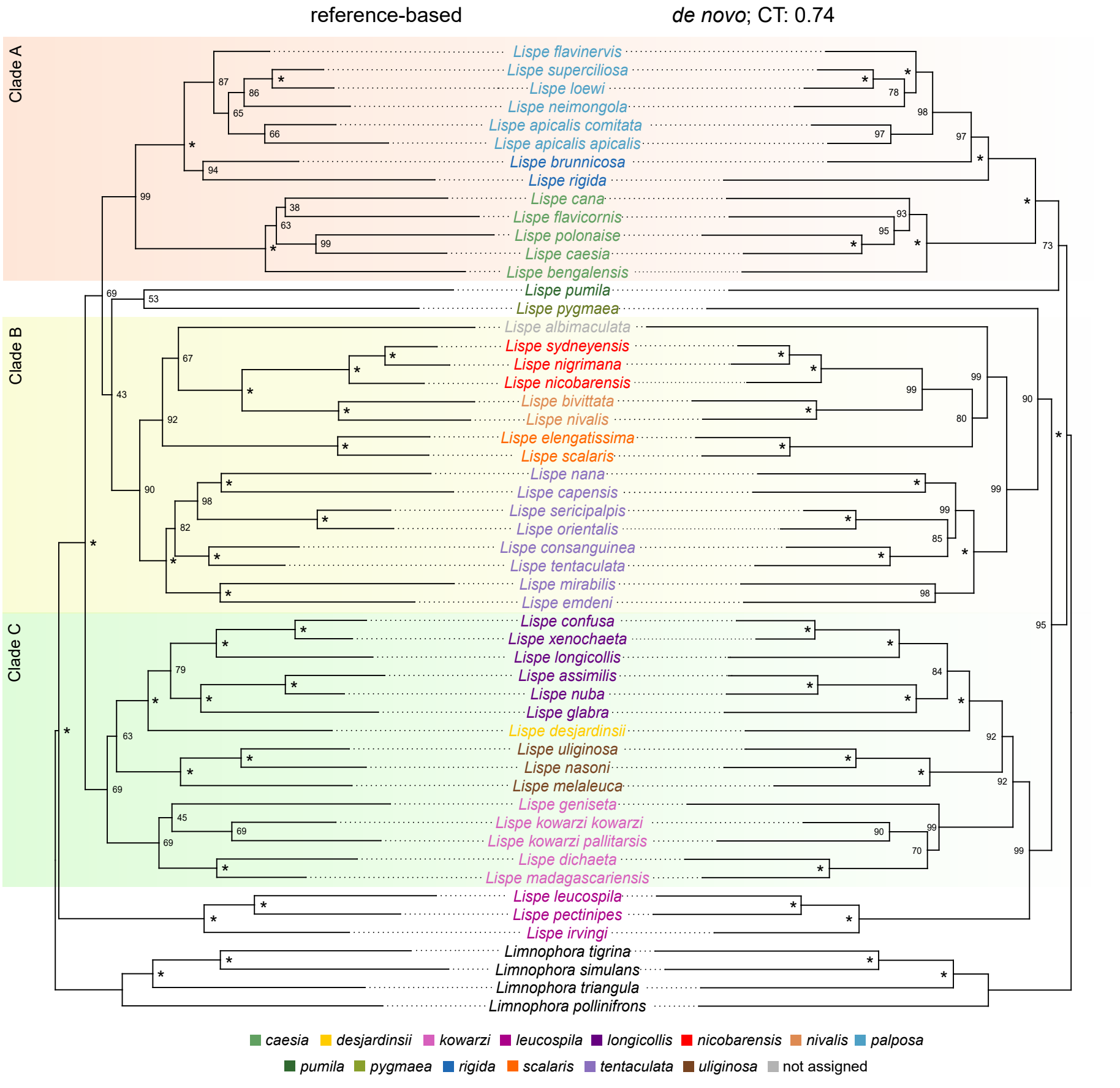
Fig. S2 Distribution of SNP numbers for datasets obtained with *de novo* assembly under different clustering thresholds (CT).

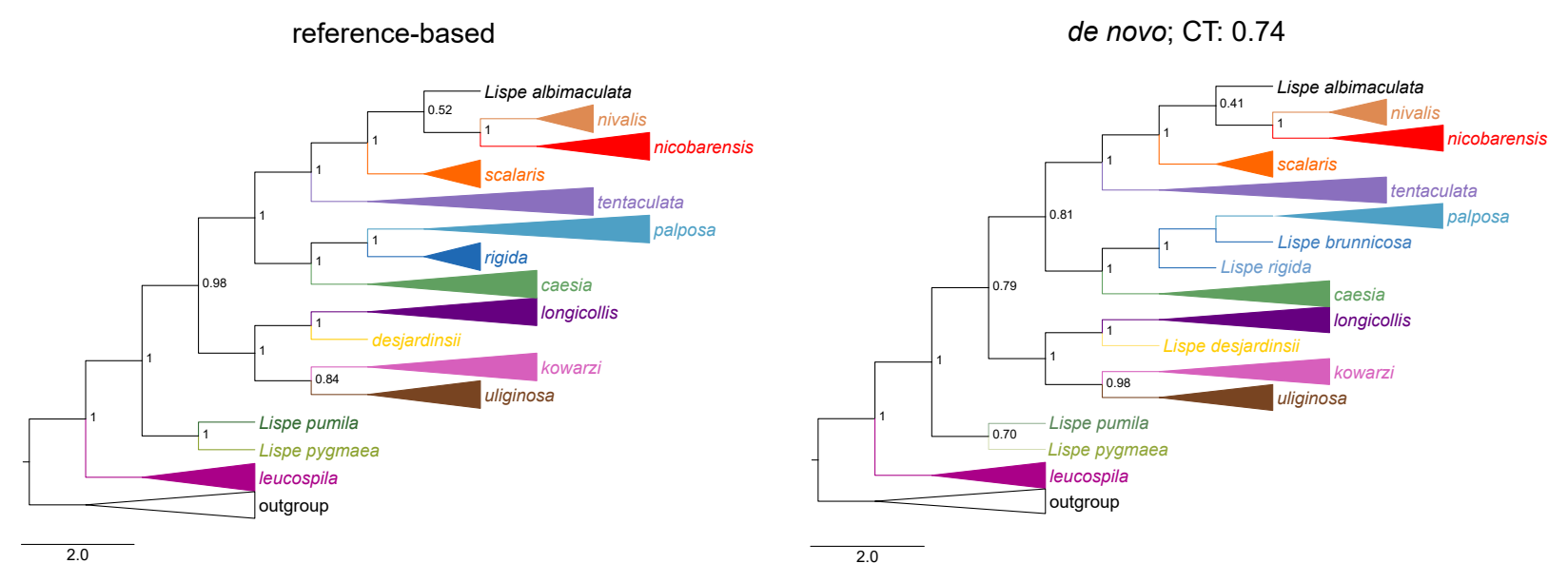
Fig. S3 Alternative topologies of phylogenetic trees obtained by *de novo* assembly under a clustering threshold of 0.75 (topology A), and a clustering threshold of 0.85 (topology B). Node support values are shown for 1000 nonparametric bootstrap replicates (BS). Species groups of *Lispe* are marked with different colours as provided. Outgroups marked with black

colour. Nodes with BS = 100 marked with an asterisk (*). Abbreviation: CT, clustering threshold.


A
B
C
D
E
F
G
H


BUSCO Assessment Results
Complete (C) and single−copy (S)
Complete (C) and duplicated (D)
Fragmented (F)
Missing (M)
CANU
C:2712 [S:1998, D:714], F:125, M:448, n:3285
FLYE
C:2958 [S:2229, D:729], F:117, M:210, n:3285
RAVEN
C:1693 [S:1671, D:22], F:205, M:1387, n:3285
SMARTdenovo
C:2000 [S:1974, D:26], F:187, M:1098, n:3285
wtdbg2
C:2542 [S:2534, D:8], F:246, M:497, n:3285
0
20
40
60
80
100
%BUSCOs


reference-based
de novo; CT: 0.74
Clade A
Clade B
Clade C
Lispe flavinervis
Lispe superciliosa
Lispe loewi
Lispe neimongola
Lispe apicalis comitata
Lispe apicalis apicalis
Lispe brunnicosa
Lispe rigida
Lispe cana
Lispe flavicornis
Lispe polonaise
Lispe caesia
Lispe bengalensis
Lispe pumila
Lispe pygmaea
Lispe albimaculata
Lispe sydneyensis
Lispe nigrimana
Lispe nicobarensis
Lispe bivittata
Lispe nivalis
Lispe elengatissima
Lispe scalaris
Lispe nana
Lispe capensis
Lispe sericipalpis
Lispe orientalis
Lispe consanguinea
Lispe tentaculata
Lispe mirabilis
Lispe emdeni
Lispe confusa
Lispe xenochaeta
Lispe longicollis
Lispe assimilis
Lispe nuba
Lispe glabra
Lispe desjardinsii
Lispe uliginosa
Lispe nasoni
Lispe melaleuca
Lispe geniseta
Lispe kowarzi kowarzi
Lispe kowarzi pallitarsis
Lispe dichaeta
Lispe madagascariensis
Lispe leucospila
Lispe pectinipes
Lispe irvingi
Limnophora tigrina
Limnophora simulans
Limnophora triangula
Limnophora pollinifrons
caesia
desjardinsii
kowarzi
leucospila
longicollis
nicobarensis
nivalis
palposa
pumila
pygmaea
rigida
scalaris
tentaculata
uliginosa
not assigned


reference-based
Lispe albimaculata
nivalis
nicobarensis
scalaris
tentaculata
palposa
rigida
caesia
longicollis
desjardinsii
kowarzi
uliginosa
Lispe pumila
Lispe pygmaea
leucospila
outgroup
0.52
1
1
1
1
0.98
1
1
1
0.84
1
1
2.0
de novo; CT: 0.74
Lispe albimaculata
nivalis
nicobarensis
scalaris
tentaculata
palposa
Lispe brunnicosa
Lispe rigida
caesia
longicollis
Lispe desjardinsii
kowarzi
uliginosa
Lispe pumila
Lispe pygmaea
leucospila
outgroup
0.41
1
1
1
0.81
1
1
0.79
1
1
0.98
1
0.70
2.0