Decontaminating genomic data for accurate species delineation and hybrid detection in the *Lasius* ant genus

Kristine Jecha[1]*, Guillaume Lavanchy[1,2]*, Tanja Schwander[1]

[1]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland.

[2]Present address: Biodiversity and Evolution Unit, Department of Biology, Lund University. Ekologihuset, Sölvegatan 37, 223 62 Lund, Sweden

*equal contribution

corresponding authors: kristine.jecha@unil.ch; guillaume.lavanchy@gmail.com

Abstract

Species identification and delineation by molecular methods has become a widely used technique and has revealed hybrids between species previously believed to be completely reproductively isolated. However, the application of molecular methods is associated with a risk of DNA contamination, which can result in the identification of false hybrids and generate inaccurate conclusions about the species' identities and characteristics. Here, we generate and analyze a dataset of nuclear SNP data and mitochondrial DNA sequences from over 1,000 *Lasius* ants to accurately investigate the species delineation and hybridization proclivity within the genus. We describe an approach, based on a combination of competitive mapping and allelic depth ratio analysis, that allows us to identify DNA contaminations and filter them from large-scale datasets. By applying this approach to the *Lasius* ants, we are able to remove interspecific contamination, and to clearly delineate each species genetically as well as identify a hybrid individual between *L. emarginatus* and *L. platythorax*.

Keywords: competitive mapping, DNA barcoding, hybridization, introgression, RAD sequencing, Formicidae

Significance

Next-generation sequencing enables large-scale studies of population structure, species delineation, and introgression for a wide range of taxa. However, the potential for cross-contamination during sample preparation is rarely considered, despite potential major impacts on the conclusions. Here, we develop a new decontamination pipeline and apply it to a large-scale genotyping dataset to study species delineation and hybridization in the ant genus *Lasius*. We find only one hybrid in a thousand individuals, a very different picture from the contaminated dataset. Testing for and removing contamination should be part of every population genomics study.

Introduction

The most widely accepted definition of a species refers to groups of populations that are reproductively isolated from others (Mayr, 1942). However, in practical applications, species distinctions rely on morphological criteria without any reproductive behavior information. Yet, even high-quality morphological identification brings its own practical challenges by being time-consuming and often requiring experienced taxonomists. These issues can be particularly apparent in sampling studies in which the identification of many difficult-to-distinguish species is required for robust analysis. The disconnect between the concept and practical implementation of the definition of a species underscores the need for a critical examination of widely used methodologies and increased implementation of more precise ones when possible.

Molecular methods of identification, often referred to as “DNA barcoding”, are used to identify distinct species based on genetic divergence at specific genetic regions (Hebert et al, 2003). The mitochondrial gene cytochrome c oxidase I (COI) has become the standard for species identification via DNA barcoding as it evolves rapidly enough to clearly distinguish between species and even diverged populations of the same species, but not too fast as to still remain reliable in grouping conspecific individuals (Cox & Hebert, 2001). However, relying solely on a mitochondrial gene can introduce limitations as it is maternally transmitted and therefore will not reveal any history of hybridization between species. Hybridization has recently become

a rich area of research as it was believed to be a very rare occurrence but is, in fact, common and widespread across many taxa (Edelman & Mallet, 2021). Analysis of nuclear DNA can not only identify hybrids but also assess the extent of introgression between the parental species within each population, providing insight into the hybridization patterns of the species pair. Yet, the use of these more advanced techniques can introduce a new problem: the risk of sequence data contamination. This can further obfuscate species delineations and lead to misinterpretations of hybrid status.

Because of the intricacies associated with methods of species identification, there is a need for novel techniques to decontaminate complex data and benchmark tests to ensure that they are precise and accurate. Investigating the ant genus *Lasius,* comprising 86 different species within the Holarctic, provides an ideal case study (Wilson, 1955; Janda et al., 2004). These species have been notoriously difficult to distinguish with many cases of cryptic species that are morphologically similar and cases of misclassifications later leading to synonymizing (Agosti & Bolton, 1990; Seifert, 1991; Seifert 1992; Wilson, 1955). Additionally, while there have been several instances of hybridization between *Lasius* species, a complete investigation into the patterns of hybridization and introgression between the various species within the genus has never been performed. A clear molecular identification technique would allow these species to be quickly distinguished and eliminate the risk of mistakenly analyzing the incorrect species.

With a dataset of over 1,000 *Lasius* ants from the Swiss canton of Vaud, we were able to examine the accuracy of a combination of morphological as well as nuclear and mitochondrial molecular identification techniques: restriction site-associated DNA sequencing (RADseq) for genotyping and COI gene analysis. In addition, we were able to use the data to further analyze these species for cryptic species and hybrid individuals. However, apparent contamination of the sequenced nuclear DNA required us to develop pipelines to decontaminate our data to varying degrees in order to ensure accurate conclusions on the evolutionary dynamics of this genus. This illustrates the necessity for rigorous quality control and decontamination to be performed when using these types of molecular techniques.

Methods

Study area and sampling

We collected samples of *Lasius* as part of the *Opération Fourmis* citizen sampling campaign (Freitag et al., 2020; Avril et al., 2019) as well as through a structured sampling process throughout the Vaud canton of Switzerland in 2019 (Szewczyk et al., 2024). About 10 workers from each colony were collected and stored in 70% EtOH until DNA extraction. In total 3,173 *Lasius* worker ants were collected, 434 from the structured sampling, and 2,739 from citizen sampling (Figure 1). These samples were identified morphologically by expert taxonomists to the species level when possible following Seifert (2018).

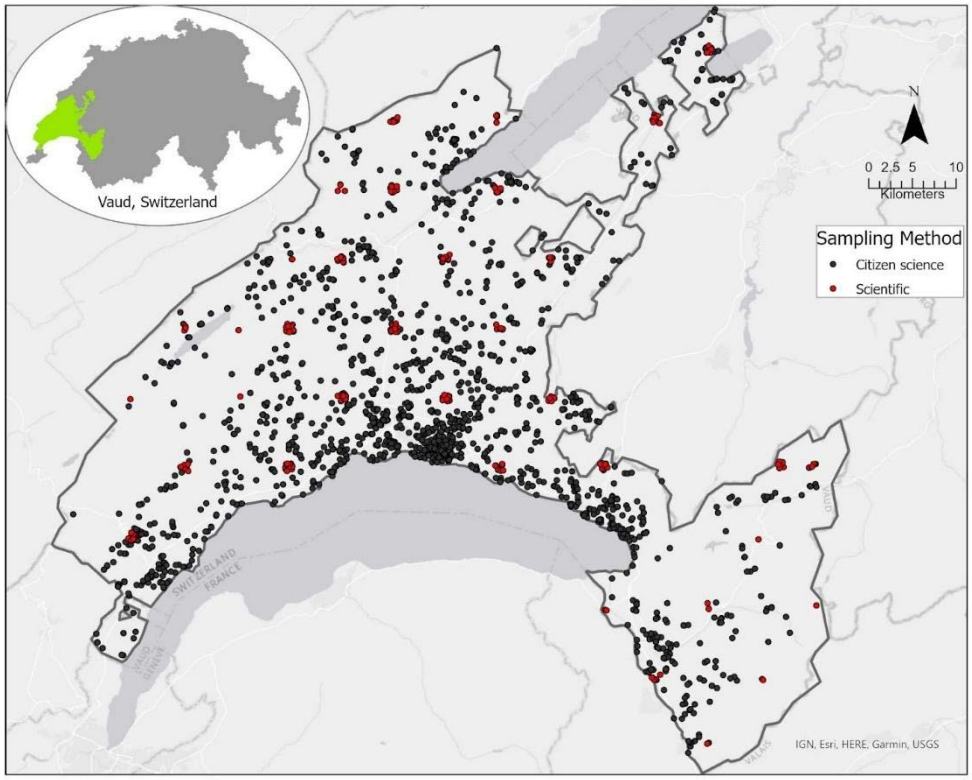

Figure 1. Map of *Lasius* colony sample locations within the canton of Vaud in Switzerland (n =3,138). Black points represent ants collected by the citizen science *Opération Fourmis* project, red points represent ants collected via structured scientific sampling. 35 samples missing geographical data and not included on the map.

DNA extraction and sequencing

For each colony, we selected one worker ant at random to undergo genetic sequencing. For each worker, three legs were used for DNA extraction. We genotyped up to 200 individuals for each (morphologically identified) species. For species that had more than 200 colonies collected in the field, we downsampled those from the citizen science campaign that were near each other in order to ensure geographically widespread sampling, resulting in a total of 1,171 genotyped individuals. The DNA extraction and COI barcoding procedures were outsourced to the Canadian Center for DNA Barcoding (CCDB) (Guelph, Canada) for 790 samples and to AllGenetics (A Coruña, Spain) for the remaining 381. A fragment of the COI gene was successfully amplified using primers LepF1 and LepR1 for 1,111 samples (Hebert et al., 2004). The specific protocols used for extraction, PCR amplification, and sequencing are available at the CCDB's website https://ccdb.ca/resources (last checked Nov 17, 2023). We generated nuclear DNA sequences using RAD sequencing protocols described in Brelsford et al. (2016). We used restriction enzymes EcoRI and MseI to digest the genomic DNA followed by ligating custom barcode adaptors to the DNA fragments, which were then amplified in 16 PCR cycles. Fragments with a size of 280 to 430 bp were selected using 2% agarose cassettes on a BluePippin (Sage Science). We sequenced these libraries on eight Illumina HiSeq 2500 lanes (single end, 150 bases) to generate the final read data.

SNP calling

We then demultiplexed the sequenced reads using stacks v2.53 process_radtags command with the options -c -q -r -t 143 (Catchen et al., 2013) and mapped the reads against a previously generated reference genome of *Lasius niger* (BioProject: PRJNA1159026) with the mem algorithm of BWA v0.7.17 (Li & Durbin, 2010). Alignment file formatting and summaries were created using samtools v1.15.1 (Danecek et al., 2021). We built loci using gstacks with options --phasing-dont-prune-hets and --ignore-pe-reads and ran populations with the option --write-single-snp to retain only one SNP per RAD locus. We were then able to filter the data with vcftools 0.1.14 (Danecek et al., 2011) keeping only genotype calls with a minimum depth of 8 and loci with a minor allele count of two which were present in at least 75% of individuals. Any samples missing more than 50% of the genotyped loci were excluded. As a quality control, we looked at the relationship between the proportion of missing genotypes and "relative

heterozygosity" (heterozygosity based on polymorphic loci) per individual. We measured relative heterozygosity as the proportion of heterozygous alleles to total genotyped SNPs.

Competitive mapping

In order to decontaminate the RADseq reads, we used a competitive mapping technique based on Feuerborn et al.'s (2020) study in which they used competitive mapping to remove modern human contaminant DNA from several ancient faunal genomic datasets. This consisted of concatenating the reference genomes of species representing potential contaminants to the target reference genome, mapping the RADseq reads to the concatenated genomes, and removing reads mapping to the non-target genomes. In our case, we combined the *Lasius niger* genome with ant genomes from seven genera to represent other ant genera also found in Vaud, Switzerland as part of the *Opération fourmis* sampling (Szewczyk et al., 2024). These included *Camponotus*, *Formica*, *Leptothorax*, *Myrmica*, *Tapinoma*, *Temnothorax*, and *Tetramorium* (Romiguier et al., 2022 and BioProjects: PRJNA445978, PRJNA819254, and PRJNA533534). We mapped the reads from each sequenced *Lasius* individual to this multi-genus concatenated genome with the mem algorithm of BWA v0.7.17 (Li & Durbin, 2010). We could then separate the reads into those preferentially mapping to the target *Lasius* genome and those that were mapping to a possible contaminant genome using R v4.2.1 (R Core Team, 2021). To filter out possible contaminants, we kept only reads that mapped either primarily or secondarily to the *Lasius* genome using the "filterbyname.sh" function of BBMap v38.63 (Bushnell, 2014). We then remapped these selected reads to the *Lasius niger* reference genome and performed SNP calling as previously stated. Any samples missing more than 50% of genotypes were excluded.

Allelic depth ratio filtering

Putative contaminations between samples from the same genus cannot be removed by our competitive mapping approach. To overcome that, we used an additional, complementary approach based on allelic depth ratio. In this paper, we define allelic depth ratio (ADR) at a heterozygous locus as the ratio of reads of the allele with most reads to the sum of all reads at the locus. It can thus range from 0.5 to <1. For a real heterozygote in a diploid individual, it is expected to be around 0.5 in all single-copy loci (Figure 2). However, if a heterozygous genotype is caused by contamination, and the amount of contaminant DNA is expected to be lower than the original sample's DNA, and ADR will be higher than 0.5. We used ADR-based filtering in two steps. First, we discarded individuals with a highly skewed ADR distribution, as

these most likely represent highly contaminated samples. We estimated the expected distribution of ADR by sampling from a binomial distribution with locus depth as the number of observations for all heterozygous sites in all individuals. We then considered that individuals whose observed median ADR (across all their heterozygous loci) was beyond the 5% percentile of this expected ADR distribution were highly contaminated and were discarded (Figure 2G).

Second, in retained individuals, we corrected the individual genotypes with ADR beyond the 25% percentile of the expected ADR distribution. We assumed that the allele with the fewest reads was derived from a contaminant and considered the individual homozygous for the other allele (Figure 2H). This filtering was performed using a custom script in R v4.2.1 (R Core Team, 2021), using the package vcfR v1.14.0 (Knaus & Grünwald, 2017).

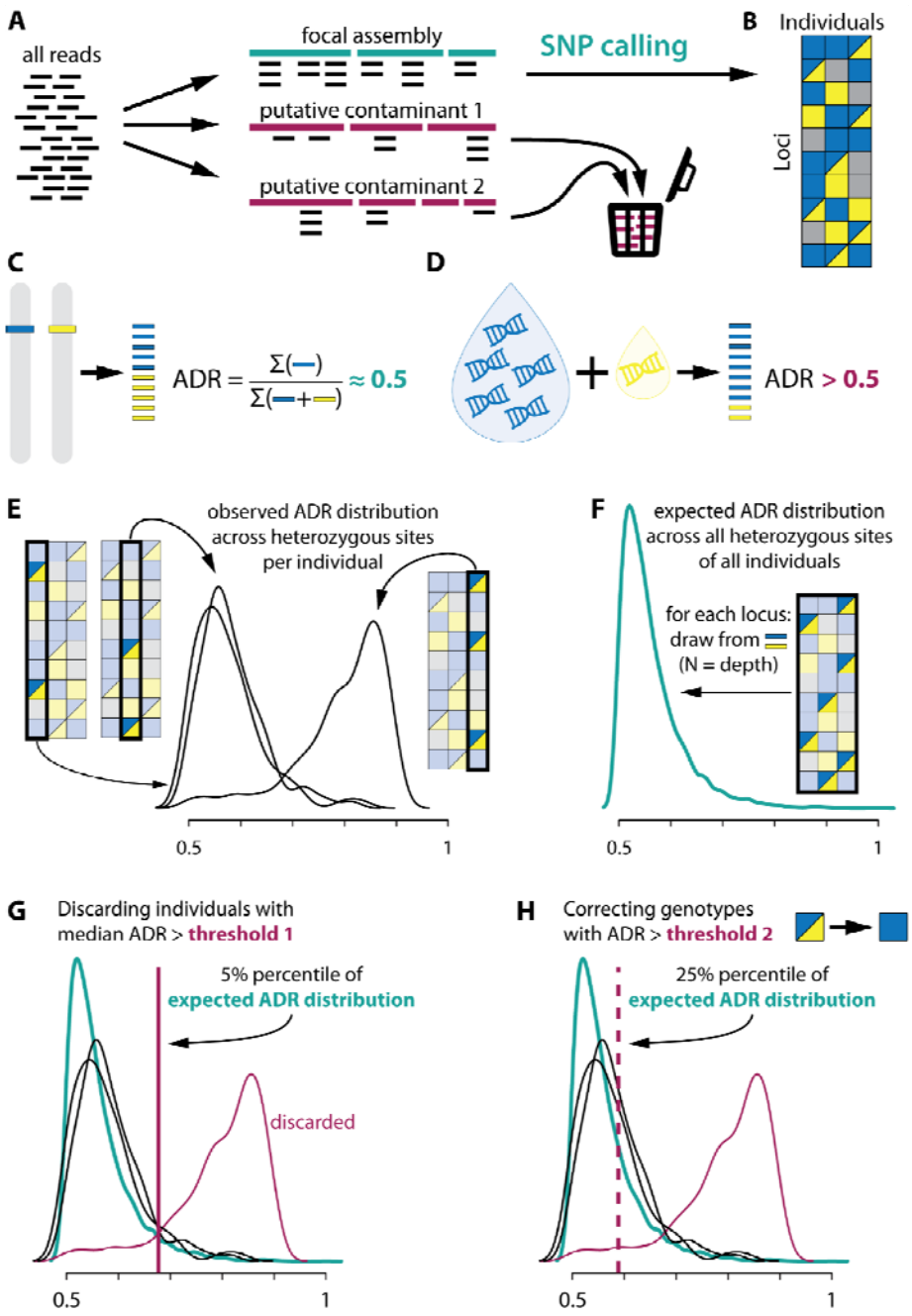

Figure 2: Overview of the decontamination pipeline. **A** Competitive Mapping. Reads are mapped simultaneously to the focal genome assembly and assemblies of putative contaminant taxa. **B** Typical dataset after SNP calling. Blue and yellow are two alleles, grey is missing data. **B-G** Allelic Depth Ratio filtering approach. **C** Allelic Depth Ratio (ADR) is expected to be around 0.5 for a real heterozygous. **D** Contaminations are expected to lead to ADR higher than 0.5. **E** We first estimate ADR distribution across all heterozygous loci for each individual independently. **F** We then estimate the expected ADR distribution across all heterozygous loci

given observed depths. **G** We discard individuals whose median ADR is beyond the top 5% percentile of the expected ADR distribution. **H** We correct genotypes with ADR beyond the top 25% percentile of the expected ADR distribution by removing the allele with the fewest reads.

Haploid allele filtering

As an alternative way to remove contaminants and assess the overall robustness of our findings, we artificially “haploidized” our competitively mapped dataset. At all heterozygous sites, we artificially removed the allele with the fewest reads. When there was the same number of reads at both alleles, one was drawn at random. We also considered homozygous sites haploid. This filtering was performed using a custom script in R v4.2.1 (R Core Team, 2021), using the package vcfR v1.14.0 (Knaus & Grünwald, 2017).

Molecular-based species identification

To investigate the congruence between the morphological identification and SNP-based identification, we used the SNP calls to create a MultiDimensional Scaling (MDS) plot with the R package MASS (Venables & Ripley, 2002) with R v4.2.1 (R Core Team, 2021). In conjunction with the MDS clustering, we ran ADMIXTURE (Alexander et al., 2009) in 10 replicates for K values ranging from 1 to 15. We then selected the best value of K and the best replicate based on Cross-Validation error. We conducted the same analysis in haploid mode with the option --haploid="*" for the “haploidized” dataset. We then visualized the results with pong (Behr et al., 2016) and chose the most representative run. In cases where the morphological identification was ambiguous or different from the genetic identification, the individual was reassigned to its genetic species identification. We were able to directly identify hybrid individuals using the ADMIXTURE data results. Any individual that contained 98.4375% or less of their main species ancestry (expected ancestry proportion after 5 generations of backcrossing) was identified as being of hybrid ancestry.

With the COI gene sequences, we created a phylogeny of the *Lasius* individuals. We first aligned the sequences using the L-INS-i algorithm of mafft v7.481 (Katoh et al., 2005). We then built a Maximum Likelihood tree using IQ-TREE v2.2.0.5 (Minh et al., 2020) using ModelFinder Plus (Kalyaanamoorthy et al., 2017) to select the best substitution model, and assessed it with 1000 UltraFast Bootstraps. Species delineation from the COI gene was compared to combined RAD-seq and morphological delineations to further confirm species identity for each individual.

Results

Decontamination of RADseq Data

RADseq genotyping was performed on 1,171 *Lasius* individuals. After excluding individuals missing at least 50% of genotypes, 981 individuals were left. The allelic depth ratio showed signs of a large amount of nuclear DNA contamination as many individuals exhibited an ADR peak around 0.9 (Figure 3A). The effect of contamination can also be seen by the dramatically increased proportion of heterozygous loci, especially in individuals that are missing a higher proportion of genotypes (Figure 3B). In order to obtain biologically meaningful results based on the SNP data created from these individuals, it was necessary to decontaminate the data.

Through competitive mapping, it appeared that the majority of the read contamination originated from ants from the *Formica* genus as the majority of non-*Lasius* reads mapped to the *Formica* genome (supplemental Figure S1). After filtering out the reads that did not preferentially map to the *Lasius* genome, an average of 64.3% (median 72.6%, standard deviation 18.1%) of reads remained. After SNP calling and discarding low-data individuals, we retained 974 samples.

The efficiency of competitive mapping as a decontamination technique can be seen clearly when comparing the decontaminated allelic depth ratio (Figure 3C) and the decontaminated proportion of heterozygous loci per individual (Figure 3D) to the contaminated data. The majority of minor allele frequencies are around 0.5, as expected.

However, competitive mapping is only able to remove contaminants outside of the *Lasius* genus, intrageneric contamination would remain. To eliminate this contamination, we compared a technique of artificial “haploidization” of the samples and an allelic depth ratio (ADR) filtering. For the latter, we discarded individuals considerably outside the expected ADR and corrected individual genotypes that were slightly beyond the expected distribution (Figure 3E and 3F). After SNP calling and removing low-data individuals, the ADR filtering retained 902 individuals.

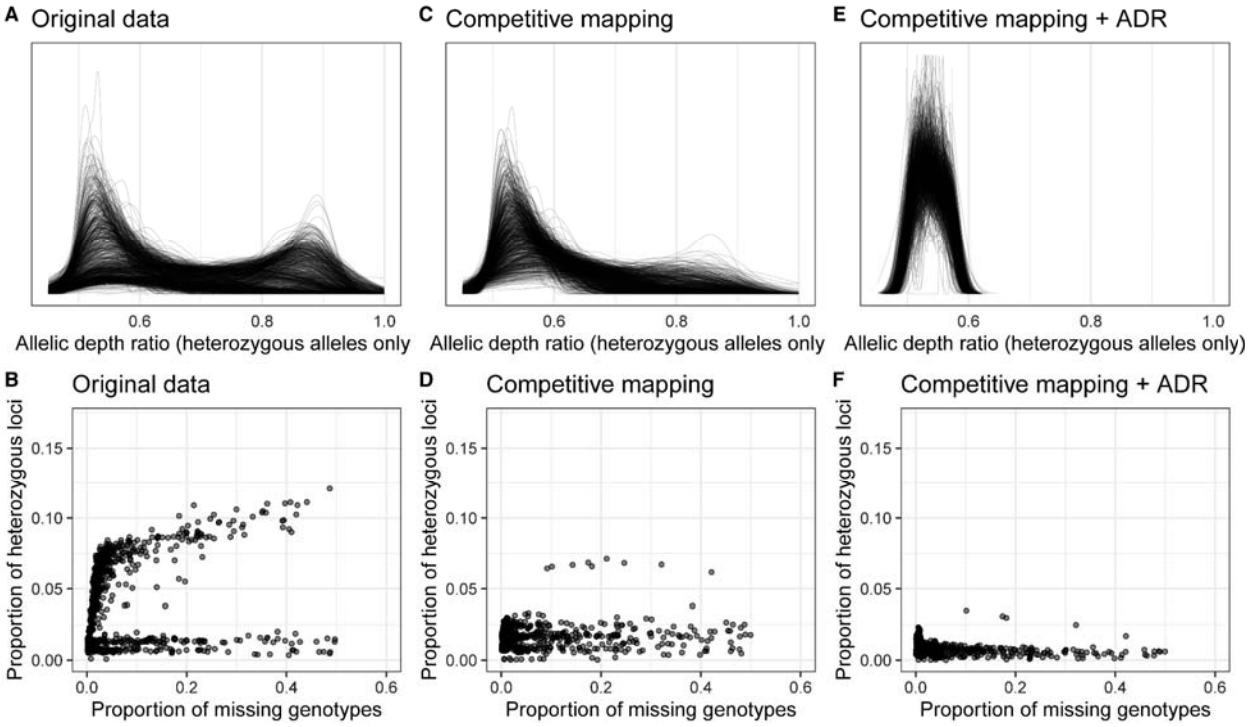

Figure 3. A) Allelic depth ratio distribution of contaminated RADseq data. B) Proportion of heterozygous loci per individual in the contaminated RADseq data. C) Allelic depth ratio distribution of competitively mapped RADseq data. D) Proportion of heterozygous loci per individual in the competitively mapped RADseq data. E) Allelic depth ratio distribution in the filtered RADseq data (filtered based on the allelic depth ratio, ADR). F) Proportion of heterozygous alleles per individual in the ADR- filtered RADseq data.

In order to identify which filtering technique was successful at eliminating the most contamination while still being useful for molecular analysis, we compared allelic depth ratio plots of each filtering technique (Figure 3). Competitive mapping alone was not sufficient to filter all contaminants, since some individuals still showed signatures of contamination (ADR peaks >> 0.5). The competitive mapping in conjunction with either ADR filtering or “haploidization” were the two best options for removing contaminants based on allelic depth frequency and MDS clustering (Figure 4 B-E). However, the “haploidization” resulted in more loss of information and would preclude additional analyses such as estimations of heterozygosity. For this reason, we selected the competitive mapping + ADR filtering method for downstream analyses.

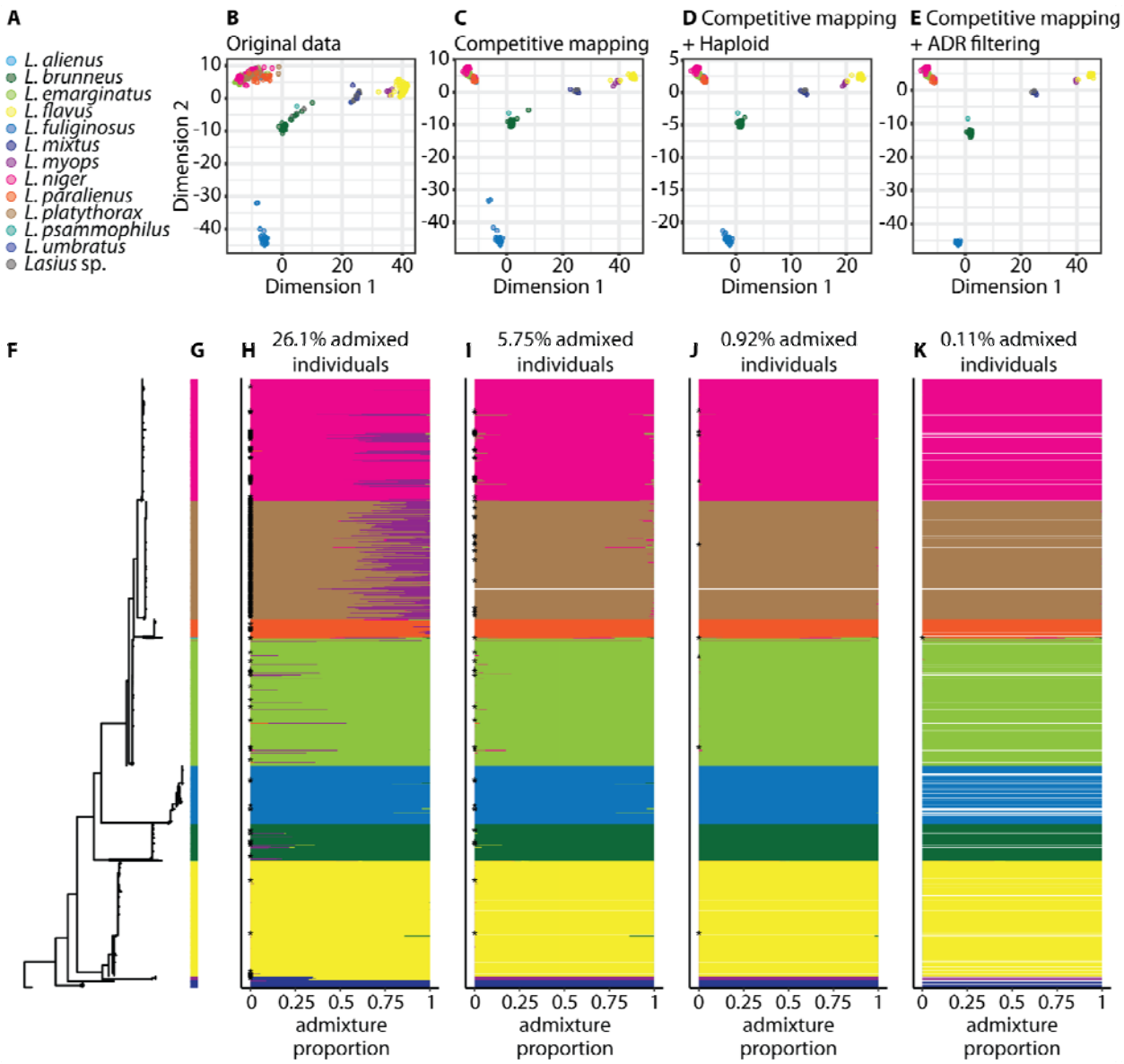

Figure 4. Taxonomic inference based on the different versions of the dataset. **A** Species color code. **B** - **E** MDS plots of original data and each decontamination process. **F** Phylogeny of the COI mitochondrial gene. **G** Final genetic-based species identification. **H - K** ADMIXTURE proportions of original data and each decontamination process, corresponding to the adjacent MDS plots. Stars denote individuals inferred to be introgressed. White lines correspond to individuals that were filtered out due to high levels of suspected contamination or low remaining SNP data. **B & H**: Original (contaminated) data. **C & I**: Competitive mapping only. **D & J:** Competitive mapping + haploidization. **E & K**: Competitive mapping + Allelic Depth Ratio filtering.

Species Delineation

Out of the 902 genotyped individuals that had sufficient data after competitive mapping and ADR filtering, 22 could not be confidently identified to the species level during initial morphological identification. The availability of nuclear genotype information in conjunction with COI sequence information allowed us to unambiguously assign these samples to a species. Additionally, 18 morphological misidentifications could be corrected based on the molecular analyses, see Supplementary Table 1 for misidentification reassignments.

Three individuals, two morphologically identified as *L. alienus* and one *L. psammophilus*, could not confidently be assigned to a genetic group in the ADMIXTURE analysis (Figure 5E) due to the low number of conspecifics. For the two *L. alienus* the ancestry proportions are unique enough that it is unlikely that they belong to any of the larger species' groups. For the *L. psammophilus* individual, in the absence of other samples of this species, it remains unclear if this is truly a *L. psammophilus* or another *Lasius* species not represented in the remaining dataset.

The combination of the MDS, COI, and ADMIXTURE analyses also allowed us to investigate the presence of cryptic species. MDS plots based on SNP distances were created for each species with at least 10 individuals to improve the resolution for detecting cryptic species (Supplementary Figure 2). However, there was no evidence of genetic subdivision within any of the species in the MDS plot, COI phylogeny, nor when the K was increased in the ADMIXTURE analysis, indicating that there are no cryptic species within the sampled *Lasius* species.

MDS, ADMIXTURE, and COI species delineations were always consistent except in one case where an individual morphologically identified as *L. platythorax* clustered with other *L. platythorax* in the MDS and had a majority of *L. platythorax* SNPs according to the ADMIXTURE analysis; however, this individual was on the *L. emarginatus* branch of the COI phylogeny (Figure 5A). The ancestry proportion of this individual is 95.06% from *L. platythorax* and 4.94% from *L. emarginatus* based on ADMIXTURE. This individual appears to have retained the mitochondria of a female *L. emarginatus* during a hybridization event in its ancestry, followed by

generations of backcrossing into the *L. platythorax* population, resulting in a predominantly *L. platythorax* nuclear DNA.

Hybrid identification

All individuals with admixture percentages of at least 1.56% (the expected proportion for a hybridization event followed by five generations of backcrossing) were classified as “admixed” (Figure 5C-E). Before filtering, 256 individuals were identified as hybrids using this cut-off. However, read filtering by competitive mapping reduced the number of hybrids to 56 individuals, artificial haploidization of the data in conjunction with competitive mapping reduced this to 9 individuals, and allelic depth filtering with competitive mapping resulted in only 1 individual classified as admixed. Indications that this single individual is a hybrid are consistent with the comparison of nuclear and mitochondrial DNA, in which the majority of this individual’s nuclear DNA is consistent with other *L. platythorax* samples; however, its COI gene appears consistent with *L. emarginatus* samples. This difference between the number of admixed individuals after the haploidization method and ADR method of filtering is because false hybrids identified in the haploidized dataset were removed during the ADR filtering due to falling above the 5% percentile of expected ADR distribution. The admixed identification of these individuals in the haploidized dataset indicates that at certain loci, more than half of the reads at that site contained an allele originating from contamination. However, as the overall allelic depth ratio frequency of these individuals was high enough for them to be excluded during the ADR filtering, this is likely a result of contamination at sites with low coverage.

Discussion

While sequencing technologies continue to advance and make genetic information more accurate and accessible, the risk of contamination in the data is rarely considered. Lack of proper quality control can easily result in drastic misinterpretations of any results. Here we show the substantial differences in biological interpretations between contaminated and decontaminated data.

We developed a method that allowed us to successfully eliminate the contamination from RADseq data of *Lasius* ants, allowing us to investigate over 900 individuals’ genetic identities. Using molecular analyses, we were able to confidently assign all individuals to a species when multiple conspecifics were present among the genetically analyzed samples. We corrected several morphological misidentifications among difficult to distinguish species (n = 18) and we were able to assign morphologically ambiguous individuals to a clear species (n = 22).

This illustrates the benefits of complementing morphological identification with molecular data to distinguish similar-looking species in an integrative way (Dayrat, 2005). In extension to this, we analyzed the known species groups to determine if there were any cryptic species. Based on our analysis of nuclear genotypes and COI divergence, we did not find evidence of cryptic species or lineages within any of the *Lasius* species in our sampling that had more than 10 individuals (*L. brunneus, L. emarginatus, L. flavus, L. fuliginosus, L. niger, L. paralienus, L. platythorax*). Those with less than 10 individuals (*L.alienus, L.mixtus, L.myops, L.psammophilus*) could not be confidently assessed due to the low sample size. Despite this, the position of the 2 *L. alienus* individuals on the COI tree assignment as a sister group to *L. paralienus* is consistent with the fact that *L. alienus* was previously distinguished as a complex of several cryptic species, two of which being *L. paralienus* and *L. psammophilus*, therefore making it likely they are all genetically close (Seifert, 1992). However, the sole *L. psammophilus* is assigned as a sister group to L. brunneus whereas previous studies have determined it to be a sister group to *L. alienus* (Janda, Folková, & Zrzavý, 2004; Steiner et al., 2004). As this was the only sample from this species, we cannot be sure if this is truly a *L. psammophilus* or is a member of a different species.

After thorough decontamination, we were able to identify only one hybrid. Interestingly, the ancestry proportion of this individual is 95.06% from *L. platythorax* and 4.94% from *L. emarginatus*; however, the COI phylogeny positions it with *L. emarginatus* indicating that this individual retains the mitochondria from *L. emarginatus*. As mitochondria can only be passed from mother to offspring, it appears that the hybridization event between these two species occurred 4 to 5 generations ago between a female *L. emarginatus* and a male *L. platythorax* and each generation produced a daughter that continued to mate with male *L. platythorax*. This is clear evidence of the hybridization between the two species and the ability to produce fertile, hybrid offspring that are capable of introgression back into their parental species' population. Previous studies have also found evidence of hybridization within the *Lasius* genus, identifying several phenotypic intermediates between *L. emarginatus* and *L. platythorax* that were proposed to be F1 hybrids between the two species (Seifert, 1992; Seifert, 2019). Hybrids and evidence of introgression between *L. niger* and *L. psammophilus* as well as between *L. niger* and *L. alienus* have also been previously recorded (Pearson, 1983; Van der Have et al., 2011). We did not encounter any cases of hybrids from these species, likely as a result of the small sample sizes of *L. psammophilus* and *L. alienus* in this survey. As our sampling did not specifically target hybrid individuals, we are able to study an unbiased sampling and observe

the frequency of hybridization between the *Lasius* species in this survey. It appears that hybridization is relatively rare among *Lasius* species. Surprisingly, we did not find any F_1 colonies, despite previous indications that they might be frequent in this genus (Feldhaar et al., 2008).

The ecological behaviors of different species can affect the frequency of hybridization between them. Several species of *Lasius* ants exhibit social parasitism which generally entails a parasitic queen invading the colony of a host species and manipulating the workers to care for her and her own brood of eggs (Janda et al., 2004). Among our sampled species, *L. mixtus* and *L. fuliginosus* exhibit temporary social parasitism in which the invading queen kills the host queen, and the host workers are temporarily mixed with the parasite's workers until the hosts eventually die off and the colony is only made of the parasitic species (Janda et al., 2004). *L. mixtus* has been observed parasitizing several species including *L. niger, L. platythorax, L. alienus* and *L. flavus* (Donisthorpe ,1922; Wilson, 1955; Collingwood, 1979; Dekoninck et al., 2004; Seifert 2018; Seifert, 1988). *L. fuliginosus* has been found to parasitize *L. niger, L. alienus, L. brunneus,* among others, and is a hyperparasite (a parasite of a parasite) of *L. mixtus* (Janda et al., 2004; Stärcke, 1944; Wilson,1955; Yamauchi & Hayashida 1968; Seifert, 1996). One would assume that these species that are often in close contact would be more likely to hybridize. However, we found no hybrids between parasite and host species suggesting the presence of barriers contributing to reproductive isolation.

The combined use of nuclear and mitochondrial DNA gives us insight into cases that would have resulted in erroneous species identification if the mitochondrial COI gene had been used alone. The *L. emarginatus* and *L. platythorax* hybrid individual would have been assigned to *L. emarginatus* as a misidentification due to its mismatched mitochondria, even though it is mostly of *L. platythorax* ancestry. This could have resulted in many misinterpretations of morphology and behavior if a mostly *L. platythorax* individual was studied believing it to be a *L. emarginatus* individual. To ensure the greatest amount of information can be applied it is necessary to adequately remove any contaminants from the data to ensure accurate analyses of both nuclear and mitochondrial sequences. As we have shown, contamination from various sources can result in the presence of false introgression in the dataset. In this case, it would have indicated widespread hybridization between various *Lasius* species that has never been recorded, as well as indicated individuals with ancestry from three species. Such "triple hybrids" have been reported in a similar dataset (Zbinden et al., 2023). Estimating the potential contribution of contaminations to this signal would be interesting.

The use of competitive mapping enabled us to exclude reads that most likely belong to other ant genera. We found that a large amount of the contamination appeared to be a result of *Formica* material being introduced into our *Lasius* samples. More precise filtering based on allelic depth frequency allowed us to further eliminate interspecific contamination within the *Lasius* genus. While this resulted in some loss of low-quality samples, the remaining samples provide a more accurate representation of the genetic makeup of *Lasius* individuals throughout Vaud, Switzerland. While we cannot be sure how this contamination was introduced into our samples, it appeared that most of the contaminated samples were those with DNA extraction performed by the CCDB (see Supplementary Table 2), highlighting the fact that even data based on wet lab work conducted by professional sequencing facilities should be quality controlled.

The ADR method we presented in this paper enabled us to drastically reduce the rate of false positive hybridization signals. It may however be slightly too conservative in the specific case where two hybridizing species differ in the amount of paralogs in their genomes. Paralogs are expected to yield signals similar to a contamination. For example, an originally monomorphic genomic region undergoing duplication, followed by divergence, could result in an AAAB signal. This would result in an ADR of 0.75, which would likely be filtered out.

RADseq allows for genotyping of non-model organisms to be done quickly and in large enough quantities for evolutionary research at the population level (Razkin et al., 2016; Rancihac et al., 2019; Herrera & Shank, 2016; Wagner et al., 2020). Alone, but especially in conjunction with COI barcoding, one can create a relatively clear picture of how populations or species intermingle with each other. Yet the presence of contamination can easily lead to incorrect conclusions. Not only can it create false-positive hybrid signals, but it may also result in other biases, such as artificially inflating heterozygosity estimates or underestimating population differentiation. Several studies have generated inconclusive or even incorrect results as a result of DNA contamination that was not fully examined before analysis (Merchant et al., 2014; Boothby et al., 2015; Koutsovoulos et al., 2016). While some of these kinds of mistakes are quickly caught, many others can easily slip through the cracks and go on to inform other studies with incorrect information. Correctly analyzing and interpreting data relies heavily on the quality of the data in the first place. Human and technical errors are unavoidable, but understanding methods in which to minimize these errors can allow for accurate conclusions to

be drawn. The method we developed here can be applied to any type of genotyping and resequencing data, and we suggest that systematically checking for the presence of contamination should be part of the default quality-check pipeline of all genomic datasets.

Acknowledgments

We thank Marc Bastardot, Marjorie Labédan, Christine La Mendola, Heather Ryder and Simon Vogel for their help processing samples. This study was supported by funding from a bequest of M. Rullens to the University of Lausanne, the fondation Herbette, the Société Vaudoise des Sciences Naturelles, and the Retraites Populaires.

Data Availability

The code used and scripts created for the data analysis can be found on GitHub: https://github.com/glavanc1/Lasius_species_ID. Demultiplexed RADseq reads can be found on SRA (Bioproject PRJNA1176511). COI sequences can be found on GenBank (accession PQ506036-PQ507160). See Supplementary Table 3 for SRA and GenBank accession numbers for each *Lasius* individual and outcome during the contamination filtering process.

References

Agosti, D., & Bolton, B. (1990). New characters to differentiate the ant genera Lasius F. and Formica L. (Hymenoptera: Formicidae). *Entomologist's Gazette*, *41*(3), 149-156.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, *19*(9), 1655-1664.

Avril, A., Dépraz, A., Schwander, T., & Freitag, A. (2019). Opération Fourmis, le premier recensement participatif des fourmis vaudoises-Contexte, méthodologie et bilan préliminaire. *Bulletin de la Société vaudoise des Sciences naturelles*, *98*, 109-120.

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., & Ramachandran, S. (2016). Pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, *32*(18), 2817-2823.

Boothby, T. C., Tenlen, J. R., Smith, F. W., Wang, J. R., Patanella, K. A., Osborne Nishimura, E., ... & Goldstein, B. (2015). Evidence for extensive horizontal gene transfer from the

draft genome of a tardigrade. Proceedings of the National Academy of Sciences, 112(52), 15976-15981.

Brelsford, A., Dufresnes, C., & Perrin, N. (2016). High-density sex-specific linkage maps of a European tree frog (Hyla arborea) identify the sex chromosome without information on offspring sex. *Heredity*, *116*(2), 177-181.

Bushnell, B. (2014). *BBMap: a fast, accurate, splice-aware aligner* (No. LBNL-7065E). Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular ecology*, *22*(11), 3124-3140.

Collingwood, C. A. (1979). The Formicidae of Fennoscandia and Denmark. Fauna Entomologica candinavica, 8.

Cox, A. J., & Hebert, P. D. (2001). Colonization, extinction, and phylogeographic patterning in a freshwater crustacean. *Molecular ecology*, *10*(2), 371-386.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*(2), giab008.

Dayrat, B. (2005). Towards integrative taxonomy. *Biological journal of the Linnean society*, *85*(3), 407-417.

Dekoninck, W., Boer, P., & Maelfait, J. P. (2004). Lasius platythorax Seifert, 1991 as a host of several Chthonolasius species, with remarks on the colony foundation of the parasites (Hymenoptera: Formicidae). *Myrmecologische Nachrichten*, *6*, 5-8.

Donisthorpe, H. (1922). The colony founding of Acanthomyops (Dendrolasius) fuliginosus Latr. *The Biological Bulletin*, *42*(4), 173-184.

Edelman, N. B., & Mallet, J. (2021). Prevalence and adaptive impact of introgression. Annual Review of Genetics, 55(1), 265-283.

Feldhaar, H., Foitzik, S., & Heinze, J. (2008). Lifelong commitment to the wrong partner: hybridization in ants. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1505), 2891-2899.

Feuerborn, T. R., Palkopoulou, E., van der Valk, T., von Seth, J., Munters, A. R., Pečnerová, P., ... & Díez-del-Molino, D. (2020). Competitive mapping allows for the identification and exclusion of human DNA contamination in ancient faunal genomic datasets. *BMC genomics*, *21*(1), 1-10.

Freitag A., Schwander T., Broennimann O. & Dépraz A. (2020). Opération Fourmis, les résultats du premier recensement participatif des espèces de fourmis vaudoises. Bulletin de la Société Vaudoise des Sciences Naturelles 99: 13-27. https://doi.org/10.5169/seals-917230

Hebert, P. D., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1512), 313-321.

Hebert, P. D., Penton, E. H., Burns, J. M., Janzen, D. H., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences*, *101*(41), 14812-14817.

Herrera, S., & Shank, T. M. (2016). RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. *Molecular phylogenetics and evolution*, *100*, 70-79.

Janda, M., Folková, D., & Zrzavý, J. (2004). Phylogeny of Lasius ants based on mitochondrial DNA and morphology, and the evolution of social parasitism in the Lasiini (Hymenoptera: Formicidae). *Molecular phylogenetics and evolution*, *33*(3), 595-614.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, *14*(6), 587-589.

Katoh, K., Kuma, K. I., Toh, H., & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research*, *33*(2), 511-518.

Knaus, B. J., & Grünwald, N. J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. *Molecular ecology resources*, *17*(1), 44-53.

Koutsovoulos, G., Kumar, S., Laetsch, D. R., Stevens, L., Daub, J., Conlon, C., ... & Blaxter, M. (2016). No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. Proceedings of the National Academy of Sciences, 113(18), 5053-5058.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589-595.

Mayr, E. (1942). Systematics and the origin of species from the viewpoint of a zoologist. *Systematics and the origin of species from the viewpoint of a zoologist.*

Merchant, S., Wood, D. E., & Salzberg, S. L. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ*, *2*, e675.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, *37*(5), 1530-1534.

Pearson, B. (1983). Hybridisation between the ant species Lasius niger and Lasius alienus: the genetic evidence. *Insectes sociaux*, *30*(4), 402-411.

R Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, 2021. *Google Scholar There is no corresponding record for this reference*.

Rancilhac, L., Goudarzi, F., Gehara, M., Hemami, M. R., Elmer, K. R., Vences, M., & Steinfarz, S. (2019). Phylogeny and species delimitation of near Eastern Neurergus newts (Salamandridae) based on genome-wide RADseq data analysis. *Molecular Phylogenetics and Evolution*, *133*, 189-197.

Razkin, O., Sonet, G., Breugelmans, K., Madeira, M. J., Gómez-Moliner, B. J., & Backeljau, T. (2016). Species limits, interspecific hybridization and phylogeny in the cryptic land snail complex Pyramidula: the power of RADseq data. *Molecular Phylogenetics and Evolution*, *101*, 267-278.

Romiguier, J., Borowiec, M. L., Weyna, A., Helleu, Q., Loire, E., La Mendola, C., ... & Keller, L. (2022). Ant phylogenomics reveals a natural selection hotspot preceding the origin of complex eusociality. *Current Biology*, *32*(13), 2942-2947.

Seifert, B. (1988). A revision of the European species of the ant subgenus Chthonolasius (Insecta, Hymenoptera, Formicidae). *Entomologische Abhandlungen*, *51*(2), 143-180.

Seifert, B. (1991). Lasius platythorax n. sp., a widespread sibling species of Lasius niger (Hymenoptera: Formicidae). *Entomologia Generalis*, *16*(1), 69-81.

Seifert, B. (1992) A taxonomic revision of the Palaearctic members of the ant subgenus Lasius s.str. (Hymenoptera: Formicidae). Abhandlungen und Berichte des Naturkundemuseums Görlitz, 66 (5), 1–67.

Seifert, B. (1996). *Ameisen: beobachten, bestimmen*. Augsburg: Naturbuch Verlag.

Seifert, B. (2018). The Ants of Central and North Europe. Lutra Verlags- und Verttriebsgesellschaft.

Seifert, B. (2019). Lasius nigroemarginatus Forel, 1874 is a F1 Hybrid between L. emarginatus (Olivier, 1792) and L. platythorax Seifert, 1991 (Hymenoptera, Formicidae). *Beiträge zur Entomologie= Contributions to Entomology*, *69*(2), 291-300.

Stärcke, A. (1944). Retouches sur quelques Fourmis d'Europe: III. Autres Lasius. *Entomologische Berichten*, *11*(255), 153-158.

Szewczyk, T. M., Lavanchy, G., Freitag, A., Depraz, A., Avril, A., Broennimann, O., ... & Schwander, T. (2024). LEVERAGING CITIZEN SCIENCE TO ASSESS RICHNESS, DIVERSITY, AND ABUNDANCE IN ANT COMMUNITIES. *bioRxiv*, 2024-06.

Van der Have, T. M., Pedersen, J. S., & Boomsma, J. J. (2011). Mating, hybridisation and introgression in Lasius ants (Hymenoptera: Formicidae). *Myrmecological News*, *15*, 109-115.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.

Wagner, F., Ott, T., Schall, M., Lautenschlager, U., Vogt, R., & Oberprieler, C. (2020). Taming the Red Bastards: Hybridisation and species delimitation in the Rhodanthemum arundanum-group (Compositae, Anthemideae). *Molecular phylogenetics and evolution*, *144*, 106702.

Wilson, E. O. (1955). A monographic revision of the ant genus Lasius. *Bulletin of the museum of Comparative Zoology, Harvard*, *113*, 2-pls.

Yamauchi, K., & Hayashida, K. (1968). Taxonomic Studies on the Genus Lasius in Hokkaido, with Ethological and Ecological Notes (Formicidae, Hymenoptera).: I. The Subgenus Dendrolasius or Jet Black Ants (With 8 Text-figures and 2 Tables). *北海道大學理學部紀要*, *16*(3), 396-412.

Zbinden, Z. D., Douglas, M. R., Chafin, T. K., & Douglas, M. E. (2023). A community genomics approach to natural hybridization. *Proceedings of the Royal Society B*, *290*(1999), 20230768.