universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Developing a pipeline for the non-destructive DNA extraction and sequencing of pinned historical insect specimens

verfasst von | submitted by

Philipp Hummer BSc MSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 831
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Zoologie
Betreut von | Supervisor:	Dipl.-Ing. Dr. Robert Kofler

Zusammenfassung

In den letzten Jahren hat sich das Sequenzieren von Exemplaren aus naturhistorischen Sammlungen – die sogenannte „Museomik“ – stark verbessert. Mit solchen Daten von 200 Jahre alten Exemplaren von *Drosophila melanogaster* konnten Scarpa et al. (2024) historische Invasionen von Transposable Elements (TEs) entdecken. Insgesamt wurden damit in *D. melanogaster* elf Invasionen in den letzten 200 Jahren gefunden, welche direkt zu einer Zunahme der Genomgröße zwischen 0.8 - 0.91% geführt haben (Pianezza et al. 2024b). So eine hohe Rate an Invasionen ist dabei aufgrund von dem, was man über den TE-Inhalt im Genom von *D. melanogaster* weiß, höchst ungewöhnlich und muss gemessen an der breiteren evolutionären Geschichte der dieser Art inflationär hoch sein. Pianezza et al. (2024b) vermuten daher, dass dies mit menschlicher Aktivität zusammenhängt. Im Besonderen könnte dies am globalen Handel liegen, da *D. melanogaster* ein menschlicher Kommensale ist und so in jüngere Vergangenheit in sekundären Kontakt mit anderen Drosophilidae gebracht wurde, wie z.B. der neotropischen *D. willistoni*. Sollte sich diese Hypothese bewahrheiten, dann würde man auch bei andere Insektenarten – besonders bei denen die kürzlich Kosmopoliten wurden – eine ähnlich angestiegene Rate von kürzlicher TE-Invasionen erwarten.

Um diese Hypothese zu überprüfen, muss man die Genome von historischen Exemplaren von unterschiedlichen Arten sequenzieren und folgend analysieren. Aus diesem Grund war es notwendig, ein Protokoll für die nicht-destruktive Extraktion von DNA zu entwickeln, welche nicht nur einen ausreichenden Ertrag erzielt, um das gesamte Genom abzudecken, sondern auch möglichst wenige physische Veränderungen bei dem Exemplar hervorruft, sodass jenes weiterhin in einer naturhistorischen Sammlung erhalten werden kann. Durch eine Vielzahl an experimentellen Extraktionen an “frischen” und historischen Exemplaren – welche sich größtenteils auf Drosophilidae und Schaben konzentrierten – habe ich ein solches Protokoll entwickelt, welches sich außerdem an die Größen unterschiedlicher Taxa anpassen lässt. Ein Novum des Protokolls liegt dabei bei dem Einsatz von “Haltestrukturen”, welche die physische Sicherheit des Exemplars während der Lysis und den Waschschritten garantieren sollen. Durch den Einsatz dieser waren die physischen Veränderungen sehr klein, was ein hohes Vertrauen in die Sicherheit des Prozederes für die Exemplare erlaubt. Die zweite Neuheit ist der Einsatz von carrier RNA, um den Ertrag bei niedrigen Konzentrationen von DNA zu erhöhen. Diese Arbeit ist dabei – soweit mir bekannt – der erste Versuch diese bei Anwendungen mit „alter DNA“ zu testen. Obwohl die erzielten Ergebnisse noch keine eindeutige Antwort liefern, sind sie bereits sehr positiv und deuten darauf hin, dass carrier RNA zukünftig Anwendung in „alter DNA“ finden wird.

Desweiteren ist es mir gelungen, hochqualitative „whole-genome“-Sequenzdaten von drei 150 Jahre alten Individuen der Deutschen Schabe *Blattella germanica* zu generieren. Aktuell konnte dies leider aufgrund der unzureichenden Qualität des aktuell besten Referenzgenoms dieser Art noch keine Antwort auf die Frage liefern, ob TE-Invasionen stattgefunden haben. Nichtsdestotrotz sind diese Daten eine wertvolle Ressource, um die jüngere evolutionäre Geschichte und Selbstdomestikation von *Blattella germanica* zu erforschen, wofür sie auch noch verwendet werden.

Abstract

During the last few years, the sequencing of specimens from natural history collections, so called "museomics", has improved considerably. Such data from 200-year-old *Drosophila melanogaster* specimens has allowed Scarpa et al. (2024) to detect historical transposable element (TE) invasions. In total, eleven TE invasions in 200 years have been discovered in *D. melanogaster*, which have directly led to an increase in genome size of 0.8 - 0.91% (Pianezza et al. 2024b). Given what is known about the TE content of the *D. melanogaster* genome, such a high rate of invasions during this time is unexpected and must be inflated relative to the species´ broader evolutionary history. Pianezza et al. (2024b) hypothesize that this is linked to human activity, specifically global trade, as *D. melanogaster* is a human commensal with a cosmopolitan range that has recently been brought into secondary contact with other drosophilids, e.g., the neotropical *D. willistoni*. If this hypothesis holds, we expect that other insect species – specifically those who have recently become cosmopolitan – may also have an inflated rate of recent TE invasions.

Testing this hypothesis requires the sequencing and subsequent analysis of the genomes of historical specimens from a range of different species. For this purpose, it was necessary to develop a protocol for non-destructive DNA extraction that not only captures enough DNA for whole genome sequencing but also causes only minimal physical changes to the specimens, so that they can continue to be preserved in natural history collections. Through many experimental extractions on "fresh" and historical insect specimens – focusing mostly on drosophilids and cockroaches – I developed such a protocol which is also adaptable to different taxa sizes. One of the core novelties of the protocol lies in the introduction of "holding structures" that ensure the specimen´s physical safety during the multiple steps of cell lysis and washing. With the use of these, I can report only minor physical changes, which allows for high confidence in the treatment being safe for the specimens. The second core novelty consists of the use of carrier RNA for assisting the retrieval of low concentrations of DNA and – as far as I am aware – this thesis represents the first attempt to test its use in ancient DNA applications. While the obtained results do not yet provide a clear answer, they are very positive and indicate that carrier RNA might be beneficial in future ancient DNA work.

Additionally, I have managed to obtain high quality whole-genome sequence data of three 150-year-old individuals of the German cockroach *Blattella germanica*. Unfortunately, this cannot yet provide a conclusive answer on whether TE invasions have occurred due to the insufficient quality of the species' current best genome assembly. Still, this data is a valuable resource for uncovering the recent evolutionary history and self-domestication of *B. germanica* and will be used as such.

Acknowledgments

University of Veterinary Medicine, Vienna

I would like to express my deepest gratitude to Robert Kofler, who allowed me to take on this project and has been incredibly supportive of me finding and testing different ideas and approaches. I would also like to express my gratitude to Sarah Saadain for our inspiring collaboration and thoughtful discussions on every aspect of this work. I warmly thank Misaki Hu for training me in lab use and DNA extraction. I am grateful to everybody in the Kofler Group, as well as the whole Population genetics institute for being supportive and having stimulating discussions. Thanks also to Maria Unterköfler for our collaboration on *Phortica* specimens.

University of Vienna

Special thanks are due to Ron Pinhasi for allowing me to use his aDNA lab resources and for his many creative ideas on different projects. I am deeply grateful to Olivia Cheronet for training me in everything related to ancient DNA lab procedures and for her insights into experimental design and the logistics of DNA sequencing. My sincere thanks go to everybody in the Pinhasi lab for sharing their experience regarding lab work. Finally, I warmly thank Mihaela Pavlicev and everybody at the Department of Evolutionary Biology for being supportive throughout this project.

Natural History Museum Vienna

I am deeply grateful to Elisabeth Haring and Martin Kapun for their support of this project and to Alex Wanka for her lab training. This work would not have been possible without Susanne Randolf, Alexssandro Camargo and Sabine Gaal-Haszler, who entrusted me to work on their valuable historical specimens.

Others

From the Smithsonian National Museum of Natural History in Washington, D.C., I would like to thank Torsten Dikow for allowing us to extract DNA from historical specimens of *Drosophila simulans* and Erin Kolski for welcoming us at the collection. I am deeply grateful to Sarah Signor and Prakash Narayanan for organizing this trip, as well as for bringing with them the required laboratory equipment which I could not have brought from Vienna. My deepest gratitude and appreciation go to my dearest Julia Mayer for her invaluable input on this text and her great insights into structuring a methodologically focused thesis. Finally, I owe gratitude to my family for their unwavering support.

Contents

Introduction

Transposable elements (TEs) are selfish genetic elements of mobile DNA sequences that replicate independently within the host genome. They were first described in Maize by Barbara McClintock in 1948, who formally published her findings in 1950 (McClintock 1950). Now, they are known to be common in eukaryotes as a whole and over the years much work has been done to describe their diversity and to classify them, such as by their transposition intermediates with the distinction into class I: retrotransposons and class II: DNA transposons (Wells & Feschotte 2020). The spread of TEs itself is highly mutagenic and is also associated with other costs such as interference with host transcription, genomic instability, and the triggering of auto-immune responses, which all make them generally very costly for an organism (Bourque et al. 2018). However, their uncontrolled spread in the genome is at some point stopped by one of several host defined defense systems and the one that is found in e.g., *Drosophila* involves piwi-interacting (pi)RNAs which control the activity of TEs (Brennecke et al. 2007). In eukaryotic genomes TEs are an omnipresent feature and the total percentage of the genome that they make up correlates strongly with genome size, yet not with complexity. In the most extreme cases, TEs can make up around 85 % of the genome (Wells & Feschotte 2020) and they also make up a vast majority of what used to be referred to as highly repetitive "junk" DNA, which by now is known to serve important roles in the organism. Further, through the process of TE gene domestication, they can also become a major source of evolutionary novelty and genes derived from them can sometimes even have essential cellular functions (Bourque et al. 2018).

However, it remains unclear how TEs are introduced into genomes in the first place. Horizontal transposon transfers (HTTs) are a rare form of horizontal gene transfer between different evolutionary lineages, which is remarkable as these are usually confined to prokaryotes. These transfers of TEs from one lineage to another – through mechanisms that are still unknown – may then trigger TE invasions (Schaak et al. 2010). Over the course of long evolutionary timescales, these transfers can have a profound impact on the evolution of eukaryote genomes such as they have had in insects (Peccoud et al. 2017). TEs have also been involved in some cases of rapid adaptation such as in the black morph of the peppered moth *Biston betularia* (van´t Hof et al. 2016) and in insecticide resistance in *Drosophila melanogaster* (Daborn et al. 2002).

The first TE invasions that were noticed were the ones of the *P*- and *I*-elements in *D. melanogaster* in 1983. Margaret G. Kidwell had noticed that the hybrids of some lineages of flies that were collected at different historic timepoints were dysgenic and that this was also dependent on parental effects. She was able to deduce that the *I*-element had started to spread sometime in

the 1930s and that the *P*-element followed in the 1950s. Eventually, both TEs spread throughout the entire species and nowadays the only unaffected individuals are lab strains established before this invasion (Kidwell 1983).

Subsequent work with different lab strains and comparative genomics has shown that two further invasions – those of the TEs *Tirant* (Schwarz et al. 2020) and *Hobo* (Periquet et al. 1989) – have occurred in the 20th century. This was then followed up by Scarpa et al. (2024) who were able to use the sequence data from 200-year-old historical specimens from Shpak et al. (2023) to identify three more TE invasions – those of *412*, *Opus* and *Blood* – that predate the establishment of the oldest lab strains of *D. melanonaster*. Further, another four invasions – those of *Spoink* (Pianezza et al. 2024a), *MLE*, *Souslik* and *transib1* (Pianezza et al. 2024b) – have been detected. These four invasions appear to have occurred after the spread of the *P-element* and together, all these discoveries add up to a total of eleven TE invasions in a mere 200 years (Fig. 1), which have directly led to an increase in genome size between 0.8 and 0.91% (Pianezza et al. 2024b). Such a high number in a short time is unexpected, as the horizontal transfers that likely trigger TE invasions have been thought to be quite rare.

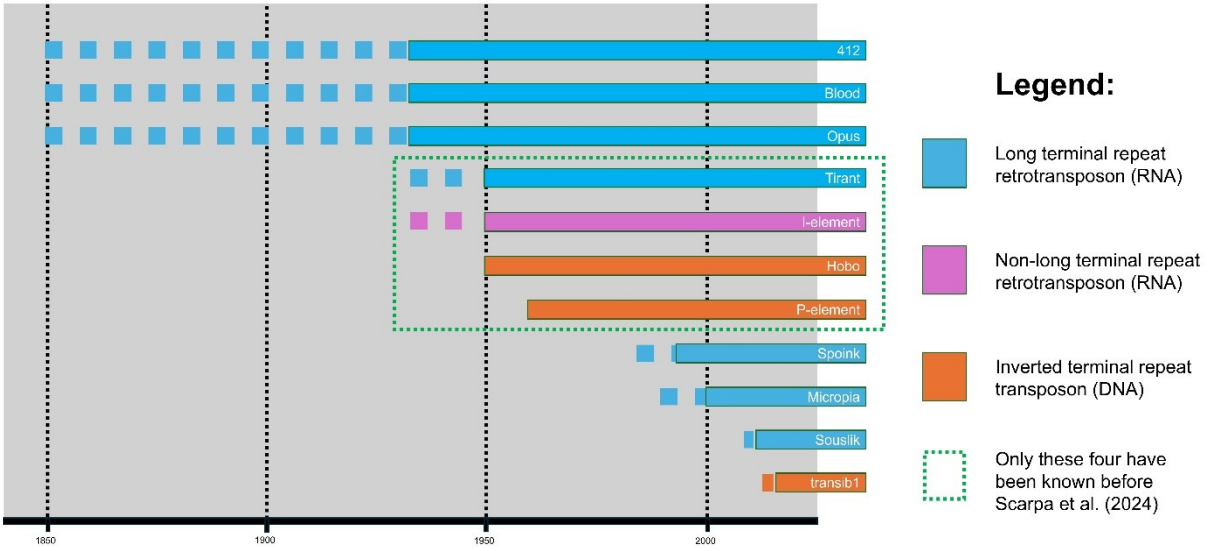

Figure 1: All known eleven TE invasions in *Drosophila melanogaster* on an approximate time scale of when they started to spread throughout natural populations. Uncertainties are visualized with dotted bars and the time frames are directly taken from Pianezza et al (2024b). Up until Scarpa et al. (2024) only four have been known.

The high rate of TE invasions in *D. melanogaster* during this time must be inflated relative to the species´ broader evolutionary history, as the total number of recognizable TE families is only around 130. This is much lower than expected, assuming a continuous rate of TE invasions of one in every twenty years. Depending on whether each family invades in multiple waves or only once, one would rather expect between 10,000 and 100,000 distinct TE families, which poses the question of why the rate of TE invasions has been elevated during the last 200 years. Pianezza et

al. (2024a&b) hypothesize that it is most likely linked to human activity, specifically global trade. *D. melanogaster* is a human commensal with a cosmopolitan range. Its environmental conditions have changed dramatically, and its habitat has increased massively, which in turn, has also led to secondary contact with other species, such as the neotropical *D. willistoni*. Human mediated habitat expansion has thus greatly expanded the chances for horizontal transfer. All of this begs the question of which species these TEs originally came from and whether other species similarly have had recent TE invasions.

The German cockroach *Blattella* germanica might represent a great model to study recent TE invasions due to its status as a pest species and its rapid spread throughout the world. A recent study by Tang et al. (2024) shows that *B. germanica* has only diverged from *Blattella asahinai* approximately 2100 years ago in human settlements around the Bay of Bengal. It then migrated both east- and westwards and was then first recorded in Denmark in 1763. By the late 19th century, it had spread throughout all of Europe and most other continents. In the early 20th century, the current distribution of *B. germanica* was shaped (Tang et al. 2019). This rapid spread can be attributed to human commensalism and advances in steam powered transportation, indoor plumbing and heating. Further, global trade has more recently also allowed for a secondary contact of the European populations with the ancestral and east-Asian ones, as well as with *B. asahinai*, which has since also become an agricultural pest in the US (Tang et al. 2024). However, not only the recency of the divergence between the two species and the speed of the subsequent spread of *B. germanica* make this species an interesting candidate to investigate its recent TE invasion history. Additionally, *B. germanica* has acquired many adaptations to human commensalism – like insecticide resistances (Daborn et al. 2002), negative phototaxis and loss of flight – in a relatively short evolutionary time scale, indicating a potential involvement of TEs.

In order to tackle questions concerning TE invasions and adaptation, the genomes of specimens that lived a long time ago can be a phenomenal resource, as they can be directly compared with those of more recent individuals to detect changes. This was done by Scarpa et al. (2024) who analyzed the twenty-five 200-year-old genomes of *D. melanogaster* that were published by Shpak et al. (2023). The sequencing of museum specimens is sometimes referred to as “museomics” and it is an approach that allows researchers to gain a deeper view of the evolutionary history of organisms by utilizing the vast wealth of vouchers that is kept in natural history collections all around the world (Fong et al. 2023). Yet, up until now, only a relatively small number of historical insect genomes are available in public databases. Sequencing and publishing genomes from historical specimens are thus very promising endeavors that could enable tackling many scientific questions, on topics such as adaptation (Shpak et al. 2023), TE invasions (Scarpa et al. 2024) and endosymbionts (Strunov et al. 2023).

However, after an organism´s death its DNA degrades and fragments over time, as the cell´s repair mechanisms are no longer active. This happens because of the exposure to water, oxygen, UV radiation or some small reactive molecules and is accelerated by temperature (Lindahl 1993). Therefore, it is expected that methodological difficulties commonly associated with ancient (a)DNA may also occur (perhaps to a lesser extent) when working with DNA from historical insect specimens. Further, a large amount of data will be needed for many types of downstream analysis, such as for using the recently released bioinformatics software GenomeDelta (Pianezza et al. 2024c), which has also been used to detect the recent invasions of the TEs *MLE*, *Souslik* and *transib1* (Pianezza et al. 2024b). It is particularly powerful because it allows to detect the absence of a TE without requiring a repeat library, meaning that it requires no prior information on the TE content of a genome. The GenomeDelta software detects systematic gaps in the alignment of the obtained short reads to a well-assembled contemporary reference genome. To ensure the detected TE candidates are statistically robust and not false positives resulting from random gaps in the alignment, a mean coverage of at least 5x is needed.

Moreover, DNA extraction from insects is usually destructive, as many protocols advise to, e.g., crush the whole insect with pestles (QIAGEN 2003; New England Biotech 2020; Asghar et al. 2015). This causes a conflict with the conservatory mission of natural history collections. If a specimen needs to be destroyed for its genome to be sequenced, the curator needs to weigh a potential scientific gain against guaranteed damage to the collection, which effectively prohibits the extraction from rare or type specimens. However, destructive extraction also limits the usability of the genetic data itself, as it is then impossible to establish a connection between the data and the intact physical specimen that remains stored in a collection. Preserving the voucher that can at any time be re-examined, can increase repeatability and help to solve questions of species identification (Mandrioli 2008). Particularly for understanding the epidemiology of vector-spread diseases, such a tight integration of a vector´s morphology and genetics may be useful (Santos et al. 2018). Thus, semi-destructive and non-destructive protocols have been developed. Semi-destructive extraction is usually performed on one or more leg(s) which may or may not be destroyed. If it remains intact, it can even be glued back on after the extraction (Patzoldi et al. 2020). Non-destructive extraction is performed by incubating the whole specimen in a lysis buffer and aims to preserve the three-dimensional structure of the outer cuticle by not digesting the chitinous exoskeleton (Phillips & Simon 1995; Gilbert et al. 2007; Santos et al. 2018). Retaining DNA from the whole individual may also be beneficial to detect/sequence transmitted pathogens (Santos et al. 2018) or endosymbionts like *Wolbachia* spp. – as has been done by Strunov et al. (2023) – or *Blattabacterium* spp.

However, both semi- and non-destructive extraction protocols will naturally yield lower concentrations of DNA than destructive methods and are influenced by the size of the specimens. Therefore, most of these approaches have often been restricted to larger species and/or target capture sequencing of high-copy barcoding regions. Considering this, as well as the previously mentioned issues with fragmentated historical DNA, it is even more remarkable that Shpak et al. (2023) successfully produced high-quality whole-genome datasets for 25 200-year-old *D. melanogaster* using non-destructive DNA extraction.

Aims of the thesis

The main goal of this thesis is to develop a pipeline for the sequencing of DNA from historic insect specimens resulting in high-quality sequencing data suitable to detect potential TE invasions. Within this endeavor, the most important task is to establish a reliable protocol for the non-destructive extraction and then purification of DNA from historic specimens from natural history collections. On the one hand, this protocol should **(1) be minimally invasive and ensure the specimens´ physical integrity**. On the other, it **(2) should result in a high concentration of DNA that ultimately enables deep sequencing covering the whole genome.** High amounts of extracted DNA are necessary to enable a deep sequencing approach resulting in reads aligning to a reference genome with a mean coverage of >5x. This in turn is required for the use of the software GenomeDelta which may then detect the presence of new TE invasions (Pianezza et al. 2024c).

The primary focus of this thesis is on documenting the development of said protocol while evaluating the physical effects of the extraction procedure on the specimens, as well as reporting on the yield, quality and fragment sizes of the DNA that results from adaptations of the protocol. Additionally, the sequence data from three individuals of *Blattella germanica* will be analyzed with GenomeDelta to allow detecting of potential TE invasions within the last 150 years. However, the sequencing of additional individuals – from a time series or multiple species – as well as a more rigorous molecular population genetics and demographic analyses are outside the scope of this thesis.

Methods

Note on DNA extraction and laboratory work

This methods chapter is intended to present the general workflow from extraction to sequencing, as well as how the sequence data for *B. germanica* was obtained. A detailed laboratory protocol for the non-destructive, minimally invasive DNA extraction from pinned insect specimens was developed as a part of this thesis. Therefore, I have decided to present it as a part of the results section, in the chapter *The final protocol for non-destructive DNA extraction*. The intent is for this chapter to serve as a complete, user-friendly, and standalone guide. It includes a safety note, a list of materials and reagents with instructions on how to prepare them, as well as a detailed step-by-step protocol for the extraction and laboratory treatment itself, along with practical tips for its application. The development of this protocol, along with the underlying rationale behind certain changes are likewise documented in the chapter *Developing a protocol for minimally invasive DNA extraction*. The aim is to provide a record of what has been learned throughout the development and offer readers an accessible entry point from which they may adapt and refine the protocol presented in this thesis.

Outline of workflow

The DNA extraction protocol was developed to specifically work on insect specimens that are pinned, which is one of the most common ways that they are preserved in natural history collections. A visual representation of the entire workflow can be found in Fig. 2. First, all paper labels were removed from the pin. To prevent potential mix-ups, pictures of the specimen with all its labels were taken. Additionally, detailed before-treatment photos were taken with a microscope (Leica M205 FA) to record the original appearance. This is useful for the general purpose of digitalization but in combination with after-treatment photos, it also serves to document any physical changes (Fig. 4-11).

First, the specimen was re-hydrated by exposing it to water steam in a box. This reverses some of the effects of the long-lasting drying process on the texture of the chitinous exoskeleton making it more elastic and therefore less brittle and prone to breakage. The size of the specimen determined which of the two holding-structure and container pairings was used (Fig. 3 B-F). After careful attachment, the individual was submerged in a buffer mix for a certain amount of time. This allowed the buffer to enter the hemocoel through gaps in the exoskeleton, such as the mouth, the anus, the spiracles and the opening that is caused by the pin (Gilbert et al. 2007). The buffer

contained the strong ionic detergent sodium dodecyl sulphate (SDS) which solubilizes lipids and proteins in the cell membrane and induces lysis of the cells in the inner tissues (Brown & Audet 2008). The buffer mix also contained Proteinase K which serves to denature the proteins in the tissue to allow access to the DNA. Yet, the buffer deliberately did not contain a chitinase to minimize damage to the outer cuticula of the specimen to minimize damage to the morphologically informative structures. Afterwards, the pinned specimen underwent multiple washing steps. Adapted from Shpak et al. (2023), it was incubated in Milli-Q water, successively increasing concentrations of ethanol and – depending on the specimen – acetone. The specimen was then "re-preparated" by adjusting positions of extremities. Then, the labels were re-attached to the pin and the specimen was dried at room temperature for a few days before being returned to its place in the collection.

The DNA present in the lysate was purified by binding to a silica column and performing several washing steps prior to DNA elution. These steps underwent a lot of experimentation and protocol refinement in order to increase quality and yield of the purified DNA, especially regarding the use of carrier RNA to help with the retrieval of low quantities of DNA. The concentrations of DNA were consistently measured using high-sensitivity Qubit for double stranded DNA and the calculated yields can be found in table A1. When these measurements were not performed shortly after the purification/library preparation/indexing, the liquid was thoroughly homogenized by up- and down pipetting beforehand. Initially, DNA purity was determined spectrophotometrically using the NanoDrop instrument by Thermo Scientific based on absorptions at 260 and 280 nm. This was proven to not be useful for aDNA. Prior sequencing on an Illumina platform, the DNA underwent library preparation. Because of age, fragmentation and damage of the DNA in our samples, we decided to perform the Santa Cruz protocol, which is designed to work on single stranded DNA (Kapp et al. 2021) and has become a standard for aDNA applications.

Afterwards, the library was double indexed with *NEBNext Q5U* and amplified non-specifically via PCR as described in Kircher et al. (2012). The appropriate number of PCR cycles was determined by running a quantitative PCR (using the *Blue S´Green qPCR 2x Mix* by Biozym), as overamplification could lead to unwanted biases. Ideally, all samples should be amplified to 1000 copies of each fragment. Usually, this takes at least 10 cycles and is intended to prevent a loss of complexity. After a purification using *NucleoMag® NGS Clean-up* (with a bead to PCR product volume ratio of 1.2x) and elution in 30 µL EBT (1 mM EDTA, 0.05% Tween-20), the fragment size distribution was checked with an Agilent TapeStation with genomic and/or D1000 ScreenTapes and the Agilent TapeStation Controller Software 4.1.1. When unwanted heteroduplexes (Orlando et al. 2021) were detected, this was corrected by performing a reamplification for 1 cycle using *Equinox® Amplification Master Mix (2X)* and the PCR product was again purified (now with a bead

to PCR product volume ratio of 1.8x). Before the proper deep sequencing run it is then advisable to perform a (screening) sequencing run with a lower yield of received data. This served as a last round of quality control and helped to assess if the samples even contained target aDNA and had sufficient quality to justify the investment of resources that us associated with deep sequencing particularly large genomes, such as that of *B. germanica*.

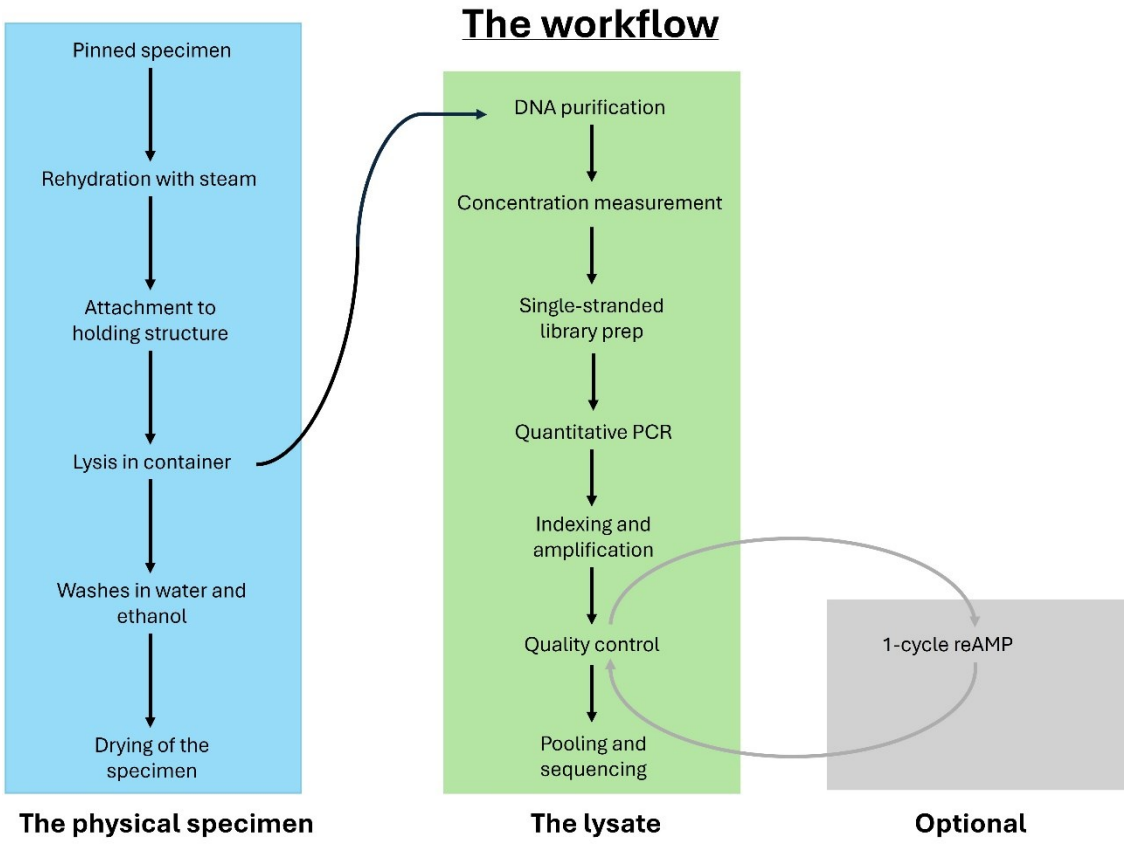

Figure 2: The entire methodological workflow of how to get sequence data from a pinned historical insect specimen.

Contamination avoidance

Since we were dealing with highly fragmented historical DNA, we avoided using regular PCR product contaminated laboratory spaces. This was also done by Gilbert et al. (2007) and Patzoldi et al. (2020) and is generally advised in Orlando et al. (2021). For this, we are collaborating with the Pinhasi group to use their aDNA clean laboratory space for the delicate laboratory procedures. This facility is commonly used to perform extraction and library preparation from ancient human DNA, which is particularly prone to contamination with the researcher´s DNA. This is both due to the low quantity of DNA from old bone material and the high genetic similarity of the samples and the researchers (Slatkin & Racimo 2016). Thus, all laboratory work is conducted under fume hoods which can be cleaned with UV lamps and the researchers themselves wear whole-body

suits and hair/beard nets. After use, all equipment and lab surfaces are cleaned with sodium hypochlorite (or DNA/RNA-ExitusPlus™) and the whole laboratory is deep cleaned monthly.

However, the standards of the aDNA laboratory space are not compatible with procedures that carry high contamination risk like our DNA extractions involving holding structures or our post-treatment care of insect specimens. Thus, we had to perform them outside of the aDNA laboratory environment. This required all necessary equipment to be brought to a well-cleaned makeshift workspace. Ideally, this equipment has never been in any room that may be PCR product contaminated. Other materials were autoclaved or cleaned with bleach and ethanol. All workspaces were thoroughly cleaned with sodium hypochlorite, 70% ethanol and water.

Sequencing of the historical cockroaches

For a first screening sequencing the samples were pooled with others from the Pinhasi group. I prepared each two libraries for the three historical *B. germanica* specimens, one for a partition that was purified using the Shpak protocol (See *Results*: *The final protocol for non-destructive DNA extraction*) and one where the Dabney protocol (See *Appendix*: *Alternative DNA purification of the lysate with the Dabney protocol*) was used. All cockroaches were submerged in lysis buffer volumes of **v** = 1500 µL and the two partitions contained 600 µL each. Each library was indexed once. Additionally, a library blank was sequenced as a negative control.

For the actual deep sequencing, I increased the number of libraries and indexed libraries by repeating the same laboratory procedures that were used for the original ones. Including these, I made 2 libraries from each of the 3 different parts of the lysate (one with Dabney, one with Shpak and one from the remaining backup lysates with Shpak – this is referred to as “Extra”) from each of the 3 specimens. The “Extra” lysate contained 130 µL for individual 1, 85 µL for individual 2 and 175 µL for individual 3. Every library is indexed a total of 3 times (= 18 indexed libraries per individual). These were pooled together such that each single individual library made up an equal proportion and sent to the Vienna Biocenter Core Facilities for paired-end sequencing. The combined pool was then sequenced on three full lanes on an Illumina NovaSeq platform. The second library blank was not yet included and will eventually be sequenced within a pool.

Sequence data polishing and analysis

From the deep sequencing we received 324 files of demultiplexed paired end sequence data (3 individuals x 3 methods x 2 libraries x 3 indexed libraries x 3 lanes x 2 for forward and reverse) in which all reads were uniformly 150 bp long due to the sequencing platform. For data polishing, a pipeline was used that was created by Sarah Saadain and that relies on the fastp command (Chen et al. 2018; Chen 2023). At first, the adapters are removed, and the forward and reverse reads are merged (parameters: --unpaired1, --unpaired2, --merge, --length_required = 15, --trim_poly_x = 5, --qualified_quality_phred = 5, --unqualified_percent_limit = 40, --n_base_limit = 5). Then the quality is polished (parameters: --disable_adapter_trimming, --qualified_quality_phred = 15, --length_required = 15, --unqualified_percent_limit = 40, --n_base_limit = 5) and the data is deduplicated (parameters: --dedup, --disable_adapter_trimming, --disable_length_filtering, --disable_quality_filtering). The quality of the data was assessed before and after each step with FastQC (Andrews 2010) and MultiQC (Ewels et al. 2016).

To find potential TE invasions in the last 150 years, the polished read data of all three individuals was merged and analyzed jointly using GenomeDelta against the currently best available genome assembly for *B. germanica* from NCBI (Genbank: GCA_000762945.2).

Methodological comparisons between Shpak and Dabney (to test the differences between purification methods) as well as between Shpak and Extra (to test the differences between lysate volumes) were performed using mixed linear models for the three metrics duplication rate (for the means between forward and reverse raw reads), mean read length (after polishing) and total received data yield in bp (also after polishing, each read length is multiplied with its frequency and this is summed up across all lengths). Before statistical testing, these metrics were summarized across the separately received data from the three lanes. This was justified as the lane was tested as a random variable but found to have no significant effect on the model. The mixed linear models account for the two methods and the random factor individual (DF = 18 – 4 = 14). The three indexed libraries per original library are treated as true replicates which is justified by them being independently amplified from different parts of the libraries. All of this was done in R (R Core Team 2022) with the use of the packages lme4 for calculating the mixed linear models (Bates et al. 2015), lmerTest for receiving *P*-values for these models (Kuznetsova et al. 2017) and dplyr for data structuring (Wickham et al. 2023). To extract the exact read length frequencies, a bash script utilizing SeqKit (Shen et al. 2016) was created, and the distribution (Fig. 13) was plotted with ggplot2 (Kassambara 2020). All used scripts are happily shared upon request.

Results

TE invasions in *B. germanica*

Running GenomeDelta for *B. germanica* with the currently best reference genome (GCA_000762945.2) and a large, concatenated file of the received data across all three 150-year-old individuals from the deep sequencing run did not detect any candidates for recent TE invasions. It is entirely possible that this is the case because there have not been any TE invasions. However, given that the used reference genome was assembled from short Illumina reads and is very fragmented (72,293 contigs and 28,065 scaffolds), it may be that repetitive sequences such as TEs are not represented in the reference genome. In that case, TEs that have recently invaded would be absent in both datasets, making them undetectable for GenomeDelta. The only way to conclusively answer the question of whether TEs have recently invaded the genomes of *B. germanica* is by running GenomeDelta with a reference genome of high quality, which can only be assembled from long reads. Such a de novo assembly is currently in the works by Sarah Saadain but it would not have been reasonable to wait for it for the submission of this thesis. For this reason, it is not yet possible to provide a conclusive answer.

Developing a protocol for minimally invasive DNA extraction

Tested specimens

The first DNA extractions were destructive and followed the QIAGEN manual (QIAGEN 2003). They were performed with many "fresh" drosophilids per sample and this amount was then reduced to a few and then to just one individual per sample. While adapting the protocol as described in the next chapters, various non-destructive extractions were performed on single recent individuals of different drosophilids (*D. immigrans*, *D. sucinea*, *D. melanogaster*), a housefly *Musca domestica*, a Colorado potato beetle *Leptinotarsa decemlineata*, two amber forest cockroaches *Ectobius vittiventris* and a brown marmorated stink bug *Halyomorpha halys*.

For further testing on historic specimens, we received a few specimens from Susanne Randolf from the Natural History Museum of Vienna: Two oriental cockroaches *Blatta orientalis*, and one common earwig *Forficula auricularia* that are all missing proper labels but were collected approximately in the 1950s. The first extractions from approximately 150-year-old material were performed with three *Phortica variegata* specimens (as a part of a cooperation with Maria Unterköfler) and three German Cockroaches *Blattella germanica*. The DNA of these three

cockroaches was the first undergoing the Santa Cruz protocol for library preparation (Kapp et al. 2021), indexing and sequencing. Later on, the final protocol has also been performed with historic specimens as old as 1868 from various drosophilids from the Natural History Museum of Vienna (*D. repleta* (10), *D. immigrans* (5), *D. histrio* (5), *D. funebris* (10), *D. buscki* (10), *D. melanogaster* (20), *P. variegata* (3) and *P. semivirgo* (3)) and with *D. simulans* (17) from the Smithsonian National Natural History Museum in Washington, D.C.

The cell lysis procedure

To lyse the cells and release DNA from the hemocoel, I adapted the protocol described by Shpak et al. (2023), along with the guidelines provided in the original QIAGEN kit manual (2003). For the lysis buffer, they describe a mixture of buffer ATL and Proteinase K, which they report to not cause significant physical damage to tested specimens. Preliminary testing resulted in consistent findings, indicating no deviations from the protocol are necessary. Other commonly used lysis/ digestion buffers – such as a potassium hydroxide (KOH) and acetic acid wash – have been reported to damage the exoskeleton (Santos et al. 2018). Duration of lysis was varied between 2 hours and 20-24 hours, depending on the specimen size and level of sclerotization. The latter was done to ensure that there is time for the buffer to enter the hemocoel and dissolve sufficient quantities of tissue. For the same reason, we also pierced holes into the ventral side of the thorax near the coxae before extracting from *L. decemlineata* and the first *E. vittiventris* (22h), like it is also described in Phillips & Simon (1995). The lysis for all non-destructive extractions was always performed at ambient temperature (20-23 °C), while the initial destructive ones were performed at 37 °C.

The preparation of holding structures

The small holding structures are based on two 1.5 mL tubes stacked into one another. The main question was then how to attach the top of a pin to the underside of a tube. It would of course be possible to drill a hole into the tube, but the first approach was using UHU Patafix to attach the pin to the tube. This proved to be a great solution that is also particularly easy and time efficient. There were also experiments with covering up the UHU Patafix with the cut-off fingers of latex gloves and parafilm. The first is not advisable because the powder on the glove may clog the pores in the silica column while the second only makes the handling slightly more difficult. However, covering up the UHU Patafix is not necessary as some experiments have shown that it is likely not

causing detectable DNA contaminations and according to the manufacturer it is entirely synthetic.

For larger specimens that would not fit in a 1.5 mL tube, a different solution was needed. The first and simplest approach was to just place the specimen in a 50 mL falcon and manipulate it using forceps via the attached pin. To prevent the specimen from floating on the surface of the liquid, a magnet was positioned beneath the falcon. This was done for the tested specimens of *M. domestica*, *L. decemlineata*, *E. vittiventris* and *H. halys*. However, there were a few issues with this approach that prevent it from being a generally good solution. The biggest issue is that during the handling the specimen will inevitably come into contact with the inside of the tube. This may lead to considerable damage to more fragile specimens. Further, the part of the pin that sticks out below the specimen may in some cases be quite long and therefore, they require the use of wastefully large volumes of lysis buffer necessary to submerge the whole specimen in it. Apart from preventable material cost, this is an issue because the resulting lysate will only have a relatively low concentration of DNA. Pins could of course be shortened as it is described in Shpak et al. (2023) for *D. melanogaster*, but this is not advisable for fragile specimens. For once, the clipping leads to inner tension in the pin resulting in a brief vibration that can damage the specimen. Furthermore, the backside of older collection boxes is often quite hard, and a clipped needle is not able to pierce it, which makes the specimen´s re-integration into the collection difficult.

Both issues were tackled for the next extraction from *F. auricularia*. First, a holding structure was built out of Lego Technic bars that could slide into the 50 mL falcons (the same that are described in *Preparation of materials*) which prevents physical contact between the specimen and the inside of the falcon. Also, a freezing bag for food was used as a container for the lysis buffer to reduce its required volume. However, the bag wrinkled which caused it to touch the specimen and for it to fall off the pin. The next idea was then to use syringe-like Combitips that were formally in use for a pipetting machine at the Pinhasi labs. These were finally tested with the specimens of *B. orientalis* and *B. germanica*, leading to the finalized version of holding structures, as they are described in the *Preparation of materials* chapter.

The DNA purification

Like the procedure for the cell lysis, the one for the purification of DNA was also adapted from Shpak et al (2023) and the used manual (QIAGEN 2003). This manual also suggests that carrier RNA (cRNA) can be added to the sample and binding buffer mix before loading the silica column. Doing so is intended to aid the retrieval of DNA when the concentration is very low. Since only low concentrations are expected for historical specimens, this would be very helpful and thus this claim was tested extensively.

Apart from the Shpak protocol, DNA was also purified with the Dabney protocol (Dabney et al. 2013), which is very commonly used in aDNA applications. This was modified by using cRNA in the same way as in the QIAGEN handbook. At first, the QiAMP MinElute Column was used but was later replaced by the regular purple MinElute Spin Columns (from the MinElute PCR purification kit) and High Pure Extender Assembly (from the Roche High Pure Viral Nucleic Acid Large Volume kit) for larger volumes. The buffer that is used to elute the DNA from the silica column during the DNA purification was varied sometimes (buffer AE: as described in QIAGEN (2003), this was used most regularly; buffer TE: 10 mM Tris-HCl, 1 mM EDTA, pH 8.0; buffer TET: buffer TE with added 0.05% Tween 20, pH 8.0).

Both tested protocols describe two rounds of elution from the silica columns during the purification process. In the experimental extractions that were aimed at establishing the protocol, these two rounds were performed with separate tubes. Measuring the concentrations showed that the first eluates were relatively stable between replicates, while the second varied strongly. Further, these second eluates also had much worse measured 260/280 absorption ratios, indicating low purity. Therefore, only the first elutes were analyzed and only their concentrations are reported. Yet, for the historical drosophilids, both rounds were eluted in the same tube to maximize the yield.

The final protocol for non-destructive DNA extraction

The following is a detailed description of how one can obtain DNA from historical specimens covering the extraction from the specimen itself, its after-treatment care and the purification of DNA. The steps from extraction to purification are based on the QIAGEN manual section on “Isolation of Genomic DNA from Tissues” (QIAGEN 2003) and its modifications by Shpak at al. (2023).

In a first step, one must decide whether to follow the protocol for "**smaller**" or "**larger**" specimens. If specimens are small enough to be submerged into 200 µL of lysis buffer in a 1.5 mL tube, follow the protocol for "**smaller**" specimens; otherwise use the protocol for "**larger**" specimens. The differences in procedure for these two sizes are explained in detail.

Safety precautions

To ensure personal safety, it must be avoided to use sodium hypochlorite (**bleach**) for the cleaning of material and surfaces during the DNA purification steps. This is because there is the possibility that it might react to dichlorine (**chlorine gas**) with the guanidine hydrochloride which is present in the buffers that are used to bind the DNA to the silicate columns (regardless of which of the protocols is used).

List of materials and reagents

Material:

- LoBind 1.5 mL tubes (for some steps they are not required to be LoBind)
- 2 mL collection tubes*
- UHU Patafix
- A hot glue gun
- A water boiler (or a microwave)
- Empty tip boxes (for the rehydration box – See Fig. 3 A)
- Standard lab equipment (pipettes, centrifuge)

Additional material for smaller specimens

- QIAGEN MinElute Spin Columns

Additional material for larger specimens

- 50 mL Combitips
- 50 mL Falcons
- High Pure Extender Assembly**
- A holding structure (e.g. one built from Lego Technic bars (See Fig. 3 F))

Reagents:

- ATL buffer*

- AL buffer*
- AW1 buffer* (prepared with ethanol in accordance with the manufacturer´s manual)
- AW2 buffer* (prepared with ethanol in accordance with the manufacturer´s manual)
- Proteinase K*
- Carrier RNA* (prepared with AE buffer in accordance with the manufacturer´s manual)
- AE buffer* (or an alternative elution buffer such as TE(T) or Milli-Q water)
- Ethanol (100%)
- Milli-Q water
- Acetone

*All marked materials and reagents have been obtained as a part of the QIAMP Micro DNA kit but may be acquired separately

**This was acquired as part of the Roche High Pure Viral Nucleic Acid Large Volume kit

Preparation of materials

The rehydration box: An empty tip box can serve as the base. In case one is only dealing with "**smaller**" specimens on short pins, it is fine to just use a box for 1000 µL tips and put some UHU Patafix over a few holes such that there is one for each of the specimens. However, in cases where the pins stand out too long above the specimens, it is better to build a rehydration box with more space. Otherwise, the closing of the lid can press the pin deeper into the hole which might damage the specimen. Such a box can be built by glueing some sterile plastic structure (e.g. cut-off inner parts from Combitips) to the inside of a large box (1000 µL tips) and then placing the tip-holding grid from a box for smaller (10 µL) tips on top of it (See Fig. 3 A)

The holding structure for "smaller" specimens: Attach a bit of UHU Patafix to the tip of a 1.5 mL tube. It should not be too little or too much (roughly half a square cell is a good amount) and needs to be molded into in elongated shape, such that the pinned specimen will be able to be fully submerged in a volume of 200 µL (See Fig. 3 B).

The holding structures and containers for "larger" specimens: For the actual lysis of the specimens, a syringe-like Combitip is needed. Separate the inner "holding structure" from the outer "container". Properly label the container and glue the tip shut with the use of a hot glue gun (See Fig. 3 C-D). Take one "holding structure", shorten the slim part by roughly 2 cm and attach some Patafix to it. Further, burn a hole through the thick part (See Fig. 3 E). This is necessary so that air can flow out of the inside, which allows the "holding structure" to compress the inside of the "container" with a shut tip. For the later washing and drying steps (and an optional soaking

step prior to the lysis), it is necessary to construct a "holding structure" which safely fits into 50 mL falcons. I did this from bleach-disinfected bars from a newly acquired "LEGO Technic 42147 Dump truck" (See Fig. 3 F). Once again, attach some UHU Patafix to it.

The washing street for "smaller" specimens: Prepare ethanol solutions of different concentrations (30, 50, 70, 80, 90 and 100%) with Milli-Q water and for each specimen pipette 200 µL volumes into 1.5 mL tubes (not necessarily LoBind). Also prepare 200 µL volumes of acetone (100%). Then place these tubes on a rack in that order, so that each specimen has its own lane.

The washing street for "larger" specimens: Prepare ethanol solutions of different concentrations (30, 50, 70, 80, 90 and 100%) with Milli-Q water in 50 mL falcons. Take an aliquot of acetone (100%) in a 50 mL falcon as well.

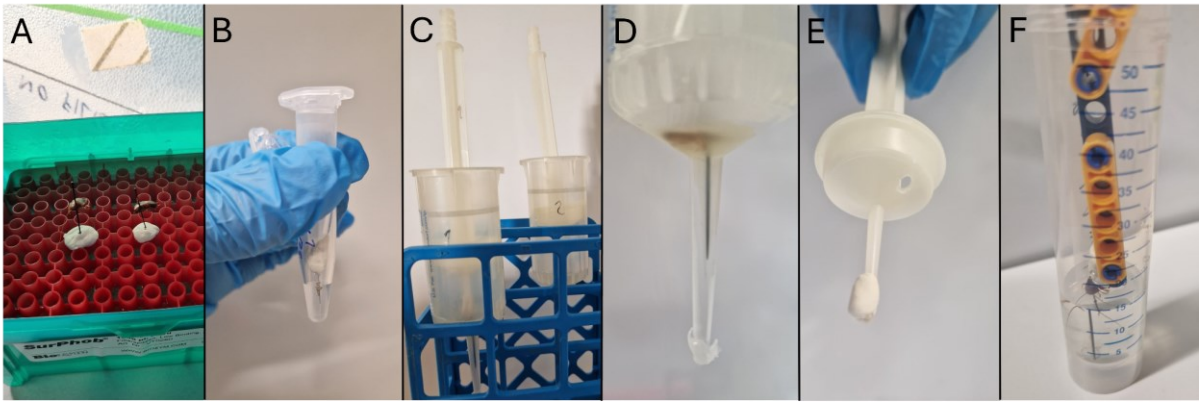

Figure 3: Material that is required and needs to be built for the protocol. **(A)** A rehydration box built from an empty micropipette tip box. **(B)** A holding structure for smaller specimens built from a second tube. **(C)** Two holding structures for larger specimens built from Combitips. **(D)** The pin beneath the insect is able to slide into the syringe-like slim tip of the modified Combitip. **(E)** A holding structure onto which larger specimens can be attached to be dipped into the lysis buffer. **(F)** A holding structure for washes of larger specimens built from Lego.

The laboratory steps for the extraction and purification of DNA

The steps for the DNA extraction itself do not require any specialized machines and can thus be performed outside of a designated molecular biological laboratory. This can be useful if one wants to extract DNA from specimens directly at the site of a historical collection or if one does not want to perform it in laboratory space that is contaminated with PCR products. The latter is of course primarily relevant for the steps up to the lysate, whereas the after-lysis treatment of the specimen(s) can be performed anywhere.

These steps can be conducted with multiple specimens in parallel, but for their physical safety it is not recommended to do it with more than ten **"smaller"** or three **"larger"** specimens at once. If the final goal is to look at the TE content across historical samples, it is recommended to order

the specimens by age. This way a cross-contamination is less likely to result in a false detection of TEs in the genome of an individual from before this TE´s horizontal transfer. For very heavily sclerotized species (e.g., *L. decemlineata*), it might be a good idea to pierce small holes into the ventral side under the Coxae prior to the treatment to allow the lysis buffer to enter through the exoskeleton.

The soaking and pre-lysis washing steps are optional. The washing is intended to remove contaminant DNA from the specimen but will likely also lead to a loss of target DNA. The soaking step aims at reducing this effect. The idea is that a thoroughly dehydrated specimen may not fully be re-hydrated by the treatment with steam in the rehydration box and therefore, it may be beneficial to rehydrate the inner tissue with lysis buffer. In cases where only minimal DNA yields are expected (small and old specimens), it might be advisable to leave these two steps out. Further, both steps require handling of the specimen which poses risk of physical damage.

It is highly advised to have a second person assisting with the after-extraction care of the specimens. This is particularly helpful if things go wrong (i.e. a specimen falls off the pin) and one is forced to deviate from the intended protocol. The second person can also re-preparate the specimens by bringing extremities back into a proper position, which for smaller specimens is only going to be possible for a few minutes after the final ethanol/acetone wash.

Pre-lysis and lysis

1. Determine the total required volume of lysis buffer: For a smaller specimen, you will need 200 µL for the lysis step and for bigger ones it may vary. As a rule of thumb, 1.5 mL should be fine for most specimens. Multiply by the number of specimens and double it in case there is going to be a soaking step. Lastly add a bit of excess volume. (E.g. for three smaller specimens with a soaking step: 200 µL volume x 3 specimen x 2 for soaking and lysis + 100 µL excess volume = 1.3 mL of total required volume)
2. Boil water and pour it into the rehydration box. Let it cool down to 80-85°C.
3. Place the pinned specimen(s) into the UHU Patafix covered holes of the grid and place it into the box. Carefully close the lid and rehydrate the specimen(s) for 20 min.
4. Prepare the needed volume of lysis buffer by mixing 9 parts **ATL Buffer** with 1 part **Proteinase K**.
5. Dispense the volumes of lysis buffer into the containers for the soaking and lysis steps. Depending on the size of the specimen this might either be a 1.5 mL LoBind tube or a Combitips Container. Make sure that they are properly labelled!

6. **Optional:** Dispense Milli-Q water into the (labelled) wash containers. For the 1.5 mL tubes ("smaller") take volumes of 200 µL and for the Combitips ("larger") take larger volumes of 10-20 mL.
7. Attach the specimen(s) to the holding structure: If you have a **"smaller" specimen**, pin it onto the holding tube with the bottom side. If you have a **"larger" specimen**, attach the top part of the pin to the Combitips holding structure by pressing the pin head into the UHU Patafix.
8. **Optional:** Submerge the specimen(s) into the soaking buffer and incubate for 5 min at RT to allow it/them to get soaked with lysis buffer: If you have a **"smaller" specimen**, simply stack the holding tube into the lysis buffer-containing tube (it might be necessary to adjust the depth of the pin in the UHU Patafix, such that the specimen is be fully submerged). If you have a **"larger" specimen**, carefully slide the Combitips holding structure into the container such that the pin below the specimen(s) fits into the slim syringe-like tip. This step is quite delicate and requires a calm hand. The specimen(s) should be completely submerged but they should not be pressed into the container´s bottom as to not damage them.
9. **Optional:** For washing, dip the specimen(s) into Milli-Q water for 5 min in the same way as in the previous step.
10. Dip the specimen(s) into the actual lysis buffer at RT for the extraction of DNA – resulting in the **lysate**. The duration is variable and depends on the species: For non-heavily sclerotized species, 2 h is a good standard regardless of size. If the specimen is heavily sclerotized (e.g., *Leptinotarsa decemlineata*), overnight lysis for around 20 h can also be justified.
11. Dip the specimen(s) in Milli-Q water for another 20 min (Keep the water as a **backup)**. Meanwhile, if you have a **"larger" specimen**, carefully transfer as much of the the **lysate** as possible from the Combitips container into a 1.5 mL LoBind tube.

After-lysis treatment of the specimen

1. If you have a **"larger" specimen**, carefully detach it from the Combitips holding structure and attach it to the other holding structure that fits the 50 mL falcons.
2. Dip the specimen into successive ethanol solutions of increasing concentrations (30%, 50%, 70%, 80%, 90%, 100%) for 5 min each. Depending on the state of the specimen(s) prior to the whole treatment, it might be better to leave out the final one or two concentrations, especially if they have noticeably shrunk in volume or if you want to extend the timeframe for re-preparation. Otherwise, the specimen becomes rigid very

quickly, as more highly concentrated ethanol dries faster. The specimen may be left in the final concentration for longer.

3. **Optional:** Dip the specimen(s) in acetone for 20 min. This leads to rapid drying and may be avoided for small specimens that are intended for re-preparation.
4. Place the pinned specimen(s) on a block of styrofoam. You or a second person may attempt to carefully re-preparate it/them by bringing their extremities back into position if that appears necessary.
5. Leave it/them to dry in the open air for three weeks.

DNA purification of the lysate (adapted from Shpak et al. (2023))

1. If you have had a **"larger" specimen**, determine the volume of lysate (**v**) that you have in µL and transfer it into a falcon. If you have had a **"smaller" specimen** it will be **v = 200**.
2. Add a volume of **v µL Buffer AL** and **v/200 µl dissolved carrier RNA** to the lysate. Close the lid and thoroughly mix by carefully turning the tube.
3. Add **v µl ethanol** (96–100%), close the lid and thoroughly mix by carefully turning the tube. Incubate for 5 min.
4. Briefly spin down the 1.5 mL tube to remove drops from the inside of the lid.
5. Carefully transfer the entire lysate to a High Pure Extender Assembly. If **v = 200**, you may instead also transfer it into a QIAGEN MinElute Spin Column (without wetting the rim). Properly label the column.
6. Close the lid, and centrifuge at maximum speed (4200 rpm) in case you are using a High Pure Extender Assembly for 4 min or at 8000 rpm in case you are using a QIAGEN MinElute Spin Column for 1 min. Make sure that the lysate has completely passed through the membrane after centrifugation. Otherwise, centrifuge again at a higher speed until the column is empty. Place the column in a clean 2 mL collection tube and discard the collection tube containing the flow-through.
7. Carefully open the column and add **500 µl Buffer AW1** without wetting the rim. Close the lid and centrifuge at 8000 rpm for 1 min. Place the column in a clean 2 mL collection tube and discard the collection tube containing the flow-through.
8. Carefully open the column and add **500 µl Buffer AW2** without wetting the rim. Close the lid and centrifuge at 8000 rpm for 1 min. Place the column in a clean 2 mL collection tube and discard the collection tube containing the flow-through.
9. Centrifuge at full speed (14,000 rpm) for 3 min to dry the membrane completely. This step is necessary since ethanol carryover into the eluate may interfere with some downstream applications.

10. Place the column in a (labelled) clean 1.5 mL microcentrifuge tube and discard the collection tube containing the flow-through.
11. **First step of elution:** Carefully open the lid of the column and apply **20 µl Buffer AE** to the center of the membrane. The volume and the elution buffer can be varied.
12. Close the lid and incubate at RT for 15 min. Centrifuge at full speed (14,000 rpm) for 1 min. You now have the **primary eluate**.
13. **Second step of elution:** If you want to have the secondary eluate in a separate tube, place the column in a clean 1.5 mL microcentrifuge tube. Carefully open the lid of the column and apply **20 µl Buffer AE** to the center of the membrane. The volume and the elution buffer can be varied.
14. Close the lid and incubate at RT for 15 min. Centrifuge at full speed (14,000 rpm) for 1 min.
15. Discard the column. You now have the **secondary eluate**. It will contain some more DNA which may either be pooled with the primary one or be kept as a backup.

Physical effects on the specimens

The before and after pictures show that the damage to the tested (non-historical) specimens (Fig 4-10) has in most cases been minimal and this applies to the historical drosophilids (Fig. 11) as well. For the tested recent dipterans – including both the drosophilids (Fig. 4-5) which fit into the small holding structure and the larger *M. domestica* (Fig. 6) which was lysed in a 50 mL falcon – only two aspects have noticeably changed: First, the treatment changes the orientation and position of extremities like legs, wings and hair. Second, it washes away the pigmentation of some ommatidia in the complex eyes. For some untreated specimens this loss of pigmentation is also present but examining historical *D. immigrans* showed that this effect is not directly caused by age. For the tested specimens of the heavily sclerotized beetle *L. decemlineata* (Fig. 7) and bug *H. halys*, the lysis only impacts the position of the legs without visibly affecting any of the informational structures of the outer cuticle. The same is true for the earwig *F. auricularia* which fell off the pin during the procedure.

Nevertheless, the same cannot be said for the two specimens of *E. vittiventris* (Fig. 8-9), which were much more fragile than all previously tested specimens. While the tagmata seem unharmed, many of their extremities broke off entirely. This is the case in both tested specimens, which only differed in the duration of the lysis step (22 versus 2 hours). As the much longer lysis time did not result in a worse physical condition, it seems that the observed damage was caused by the handling during the procedure rather than by the exposure to the lysis buffer itself. This indicates

that fragile specimens such as cockroaches are thus not suitable for extraction without the use of a holding structure. However, there was much less damage for the two specimens of *B. orientalis* when the finalized holding structure was used (on one, the tarsi of the left hind leg broke off (Fig. 10)). For the three specimens of *B. germanica*, the ootheca fell off from one specimen but was able to be glued back on. Other than that, no visible damage was observed.

Even with the use of a holding structure, non-destructive DNA extraction from fragile historical specimens is a delicate procedure and damaging the specimen remains possible. Yet, during the extractions of the in total 86 historical drosophilid specimens (including the nine *Phortica* sp.), there were only very few complications. Across all *Drosophila* sp., five specimens fell off their pins and had to be re-pinned. Further the heads broke off from two specimens, once during re-pinning. During the extractions of the nine specimens of *Phortica* sp., which were all pinned with pins that were too long to allow for the use of the small holding structure, the head broke off one specimen and for one both the head and one of the wings broke off. In all cases the heads were able to be glued back on. Losses of the antennae or (parts of) legs were not generally recorded, but the former was noticed to have happened at least once for a specimen of *D. funebris* from 1887 (Fig. 11).

Drosophila immigrans (20.0x)

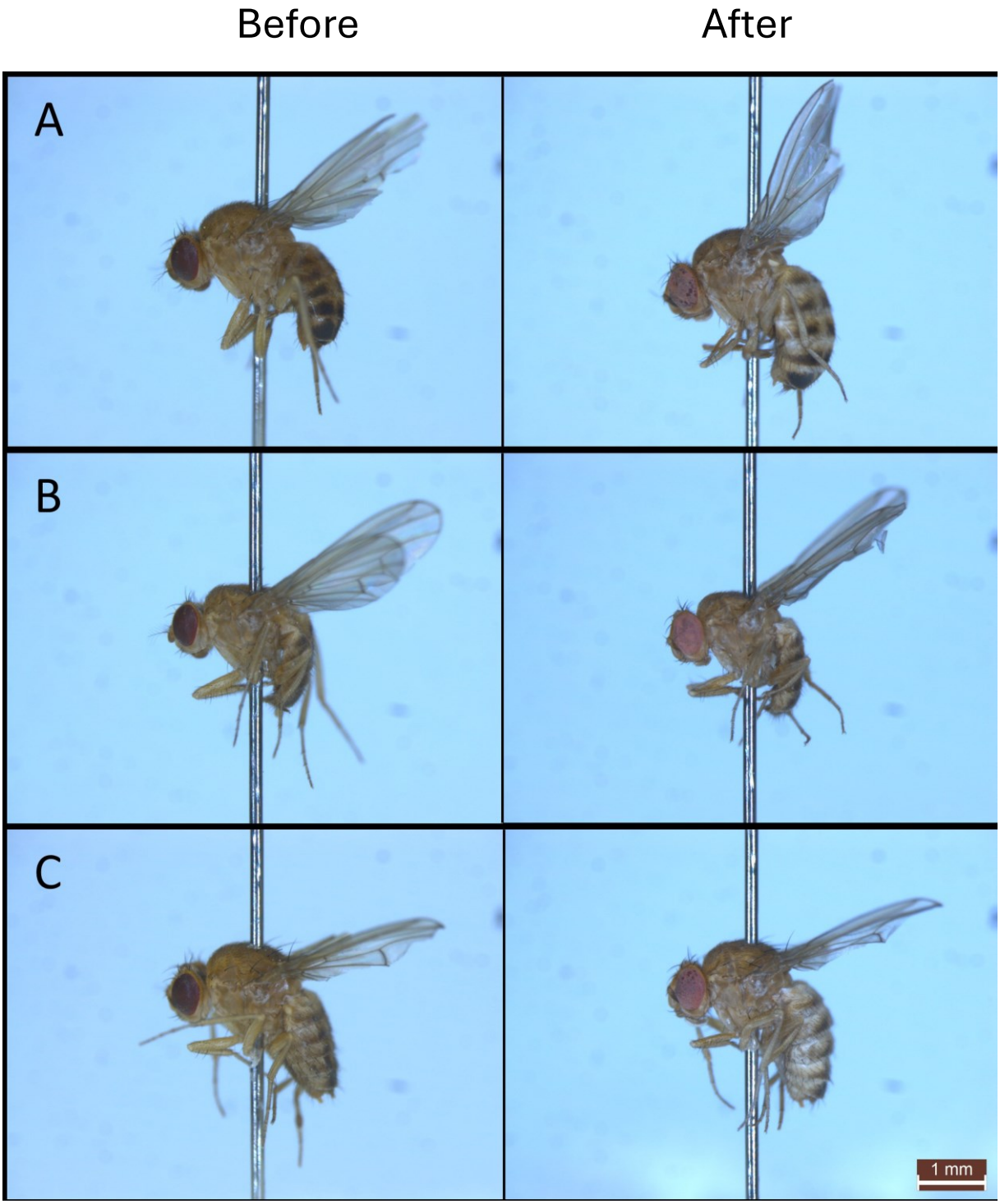

Figure 4: Pictures from the lateral view of *Drosophila immigrans* before (left side) and after (right side) the treatment. A-C show three different specimens. The loss of pigmentation in the eyes is very apparent. The magnification is 20.0x.

Drosophila sucinea (20.0x)

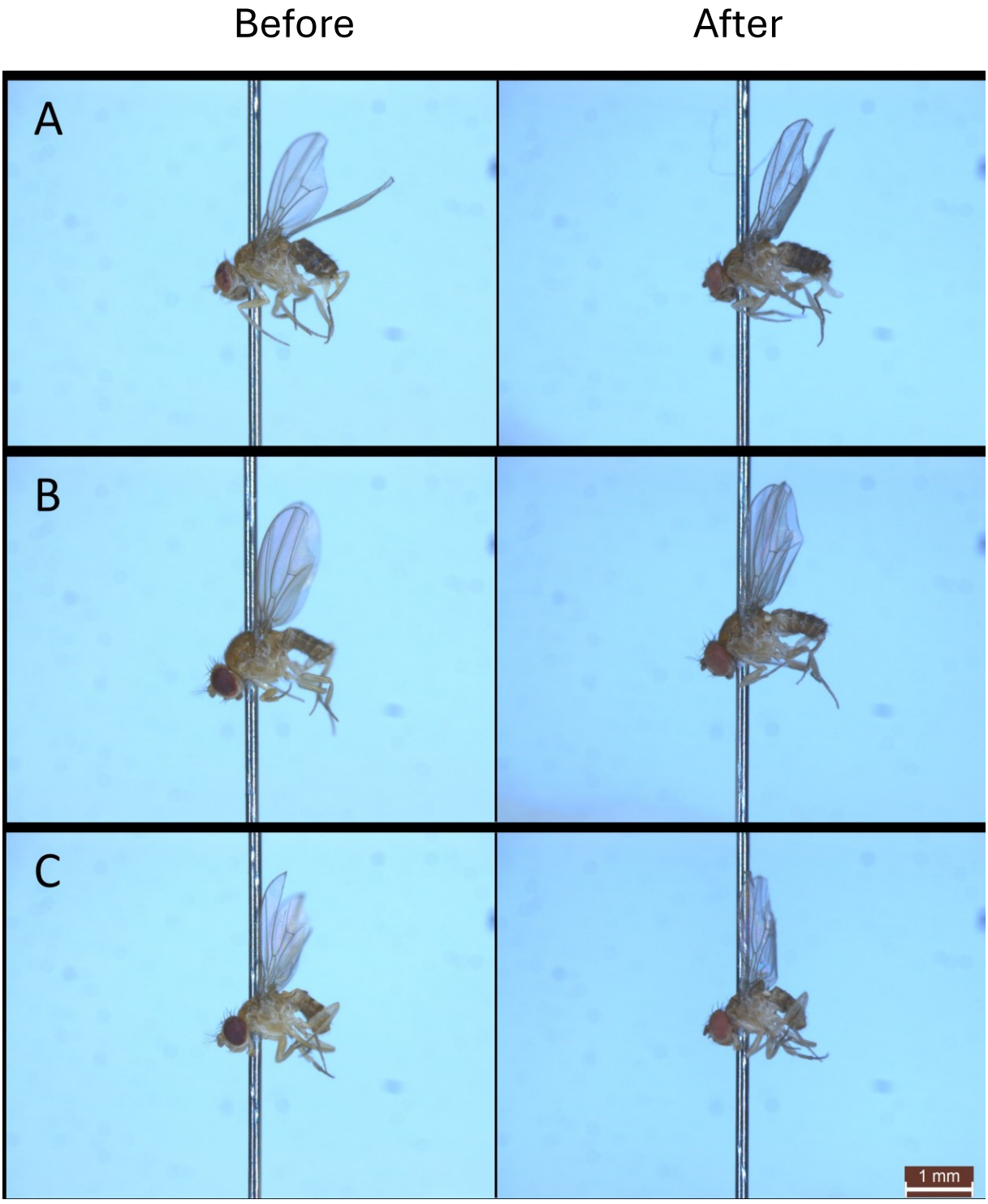

Figure 5: Pictures from the lateral view of *Drosophila sucinea* before (left side) and after (right side) the treatment. A-C show three different specimens. In specimens shown in B and C, the loss of pigmentation in the eyes is as apparent as in Fig. 4. However, it appears that in the specimen shown in A the loss of pigmentation has mostly been completed prior to the extraction. The magnification is 20.0x.

Musca domestica (10.0x)

Before After

Figure 6: Pictures from the lateral view of a housefly (*Musca domestica*) specimen before (left side) and after (right side) the treatment. In A and B show different angles of the same individual. It appears as if the pigmentation of the eyes changed similarly to Fig. 4 & 5. The magnification is 10.0x.

Leptinotarsa decemlineata (10.0x)

Before After

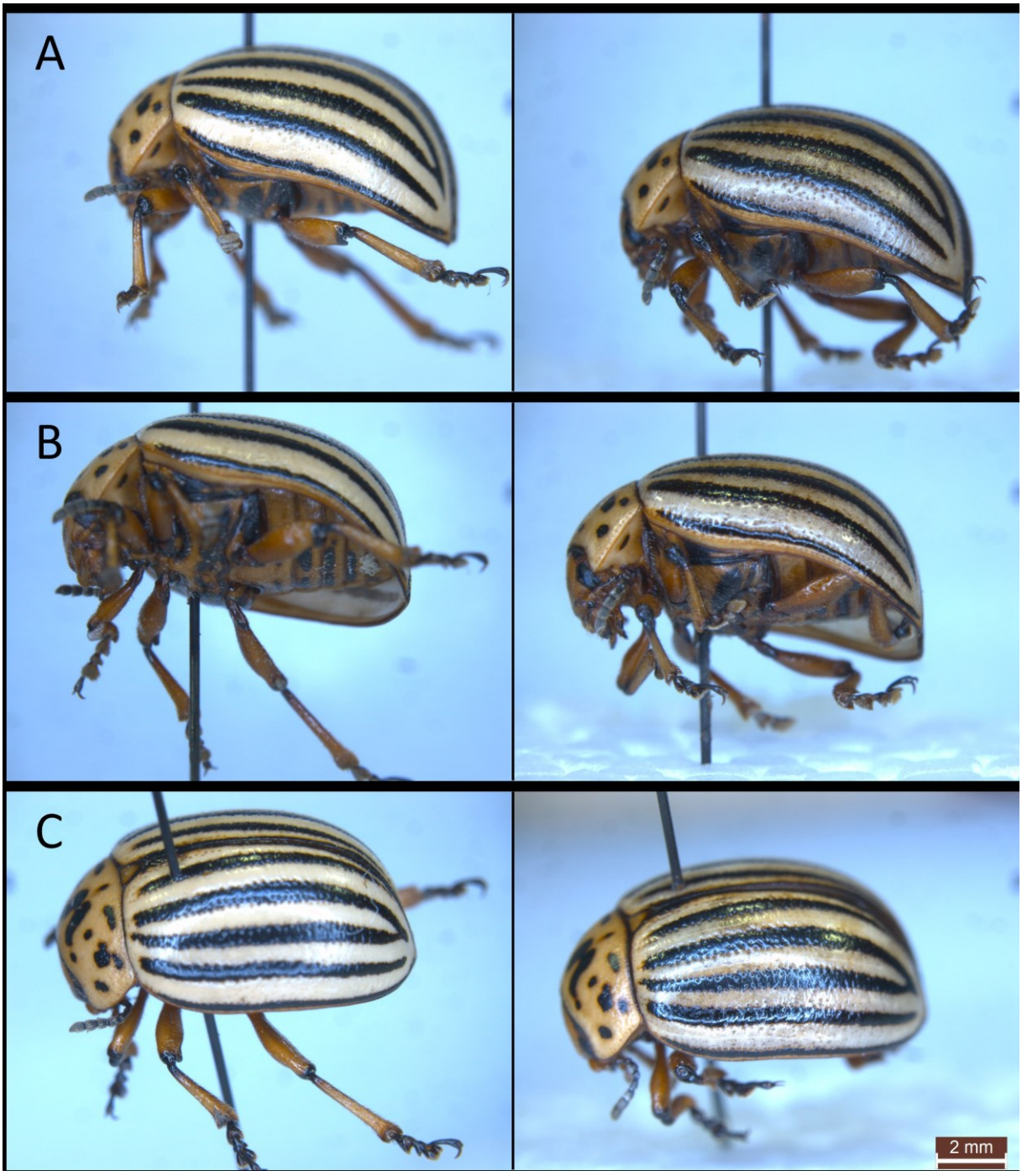

Figure 7: Pictures from the lateral view of a Colorado potato bug (*Leptinotarsa decemlineata*) specimen before (left side) and after (right side) the treatment. A-C show different angles of the same individual. The magnification is 10.0x.

Ectobius vittiventris – 22h lysis (7.5x)

Before After

Figure 8: Pictures from the dorsal (A) and lateral (B) view of a *Ectobius vittiventris* specimen before (left side) and after (right side) the treatment with a 22h-lysis. The antennae and various segments of the legs have fallen off through the treatment and the abdomen has considerably lost in volume. The magnification is the microscope's technical minimum of 7.5x.

Ectobius vittiventris – 2h lysis (7.5x)

Before After

Figure 9: Pictures from the dorsal (A) and lateral (B) view of a *Ectobius vittiventris* specimen before (left side) and after (right side) the treatment with a 2h-lysis. The antennae and various segments of the legs have fallen off through the treatment. There is little difference between the after-treatments conditions of this specimen after 2h and the one in Fig. 8 after 22h. The most noticeable difference is that in this one the abdomen has kept more of its original volume. The specimen detached from the pin and was re-pinned through the original perforation resulting in a looser fit. The magnification is the microscope's technical minimum of 7.5x.

Blatta orientalis – (7.5x)

Before After

Figure 10: Pictures from the lateral view of a *Blatta orientalis* specimen before (left side) and after (right side) the treatment. The tarsi of the left hind leg broke off. The use of the finalized holding structure thus led to only minimal damage in comparison to the two *Ectobius vittiventris* specimens (Fig. 8 & 9). The magnification is the microscope's technical minimum of 7.5x.

Drosophila funebris – specimen from 1887(4x)

Before After

Figure 11: Pictures from the lateral view of a *Drosophila funebris* specimen (specimen number 5) from 1887 before (left side) and after (right side) the treatment. The antennae broke off and much like the other dipterans (Fig. 4-6) the pigmentation of the eyes changed. Photos were taken by Sarah Saadain with a stacking microscope at the Natural History Museum of Vienna. The magnification is 4.0x.

The yields of extracted DNA

All measured concentrations of DNA in the first or combined first and second eluates, as well as the calculated yields from all tested species along with noteworthy details regarding the extraction, can be found in table A1 as a part of the supplementary material. It is important to note that not the entire yield is ever accessible to sequencing. Each of the performed measurements (Qubit, NanoDrop and TapeStation) requires volumes of 1 µL and the entire input volume may not always be able to pass through the column.

Regardless of species, all extractions with the finalized protocol have successfully resulted in measurable concentrations of DNA. This has also generally been true for most extractions with earlier versions of the protocol. Only one whole round of extractions with the then newly tested Dabney protocol (Dabney et al. 2013) ever failed completely and did not result in measurable concentrations of DNA, which was likely due to human error. However, later experiments aiming at uncovering the source of an apparent systematic contamination in the extraction blanks of historic drosophilids (serving as a negative control) revealed that cRNA itself is detectable with High Sensitivity Qubit for double stranded DNA. Across the 12 extraction blanks the measured yield ranged from being out of measurable range (for practical reasons this is considered as 0) to 60 ng (mean = 20.67; median = 13.26). It is possible that cRNA improves the capture of actual contaminations which would otherwise not be detectable. However, if one assumes that these blanks are free from significant amounts of actual contamination and that the measured concentrations of DNA are subsequently caused by cRNA, then 1 µg of applied cRNA during the QIAGEN DNA purification (QIAGEN 2003) may lead to a considerably high false signal. Therefore, the measured yields cannot fully be trusted and should not be taken at face value, as a large but unknown fraction of it is likely caused by cRNA.

Some or even many of the samples from historic drosophilid samples might not have resulted in eluates with a measurable quantity of actual DNA. If one simply subtracts the mean blank yield from the measured sample yield, then only 33 of 86 samples (38.4%) have a positive total. If the median is used instead, which might be more conservative considering there are only a few very large blank yields, it is instead 49 (57%). However, this does not mean that we can only expect to retrieve sequence data for roughly every other specimen. For once, Hamid R. Ghanavi who performed the DNA extractions in Shpak et al. (2023) without the use of cRNA revealed to us that they have never been able to directly measure the DNA concentration in any of the eluates of their 25 200-year-old *D. melanogaster* specimens. Regardless of this, they have been able to produce high quality WGS data that was sufficient for an analysis with GenomeDelta (Pianezza et al. 2024b; Scarpa et al. 2024). Further, it is hypothesized that the way cRNA improves yields is by non-

specifically binding to the silica matrix and tube wall instead of the sample DNA (Shaw et al. 2009; Sundari et al. 2021) and the presence of low quantities of sample DNA would not necessarily noticeably increase the measured yield.

Considering that cRNA is detectable with the employed method of DNA concentration measurement and that the measured yields it leads to are highly variable, deep analysis of these obtained yields would be inconclusive. However, some overall tendencies can still be observed. As cRNA is detectable, the introduction of cRNA has greatly increased the measured yields in all examined cases. The size of the specimen is also a clear factor, with the drosophilids leading to lower measured yields than the much larger other tested specimens. Within the drosophilids, extractions from the larger specimens of *Phortica* sp. and *D. immigrans* also resulted in higher yields. Another important factor is age, and it would be expected that there is a negative relationship due to the instability of DNA. However, I decided not to investigate this because of the unreliability of the measurements. Patzoldi et al. (2020) tested the relationship between age and DNA yield and found no significant correlation. Instead, they argue that the yield is mostly determined by the immediate postmortem treatment of the specimen.

Sequence data

From the deep sequencing we received one compressed folder for each of the three lanes, which added up to a size of 466.56 Gb. After decompression, these folders contained paired end sequence data that was already demultiplexed into 324 files (54 samples x 3 lanes x 2 for forward and reverse). These raw files contain on average 17.20 million reads (s^2 = 18.53 million), the mean GC content was 60.0% (s^2 = 6.9) and the mean duplicate rate was 20.3% (s^2 = 21.3). After adapter removal, quality trimming and de-duplication, the number of reads across the 162 merged (forward and reverse) files has been more than halved to an average of 8.23 million (s^2 =7.27 million). Yet, this stringent polishing has also resulted in a mean GC content of 34.6% (s^2 = 0.5) and duplication rate of 0.28% (s^2 = 0.04) which both indicate high quality (Fig. 12).

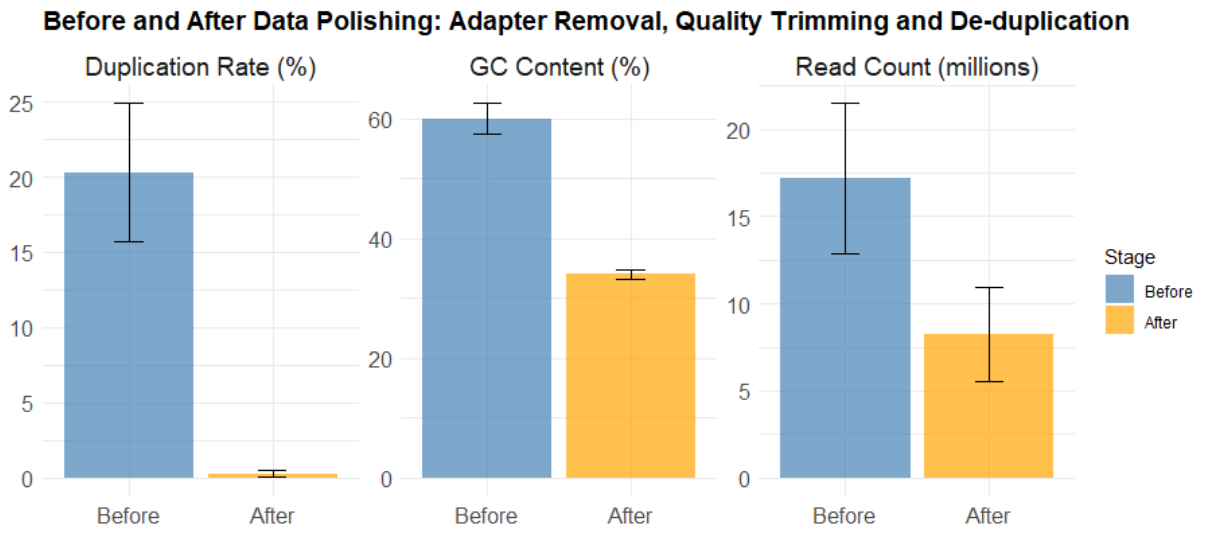

Figure 12: Three different scores – duplication rate (%), GC content (%) and read count (millions) – before and after adapter removal, quality trimming and de-duplication. The stringent data polishing has removed approximately half of all reads yet has been successful in achieving a good quality.

The Shpak and Dabney protocols

When directly compared with different partitions of the same three cockroaches, the Shpak protocol resulted in higher measured yields than the Dabney protocol. However, it seems likely that this difference is caused by cRNA. In both cases, 1 µg (1 µL with a concentration of 1 µg/µL) was added to the mix of lysate and binding buffer. However, for every 200 µL of lysate this mix was 600 µL for the Shpak protocol, while it was 2800 µL for the Dabney protocol. Thus, the concentrations were very different (1.66 ng/µL and 0.36 ng/µL). Therefore, it cannot really be concluded that the reagents in the QIAGEN protocol are responsible for an increase in eluate yields by themselves.

When using mixed linear models for comparing the files that were received from the two different method groups – Dabney and Shpak (excluding "Extra") – we can see some differences between them. For once, the Shpak protocol decreases the duplication rate by 3.04% ($P = 0.02934$) which indicates that it is slightly better at retrieving complexity. Also, it produces (post-polishing merged paired-end) reads that are on average 5.28 bp shorter ($P < 0.001$), which can also be seen in the read length distribution (See Fig. 13). The samples that were purified with the Shpak protocol – and which thus had a higher concentration of cRNA – have their peak at a shorter sequence length. cRNA may thus improve the retrieval of short fragments of DNA, which are known to be lost during the laboratory procedures prior to sequencing (Glocke & Meyer 2017). Yet, the Shpak protocol also decreases the total yield of polished data (each length multiplied by its frequency,

summed up across all lenghts) substantially by 34.36% ($P < 0.001$) relative to the mean that was obtained from Dabney (447,593,273 bp across all 18 samples x 3 lanes).

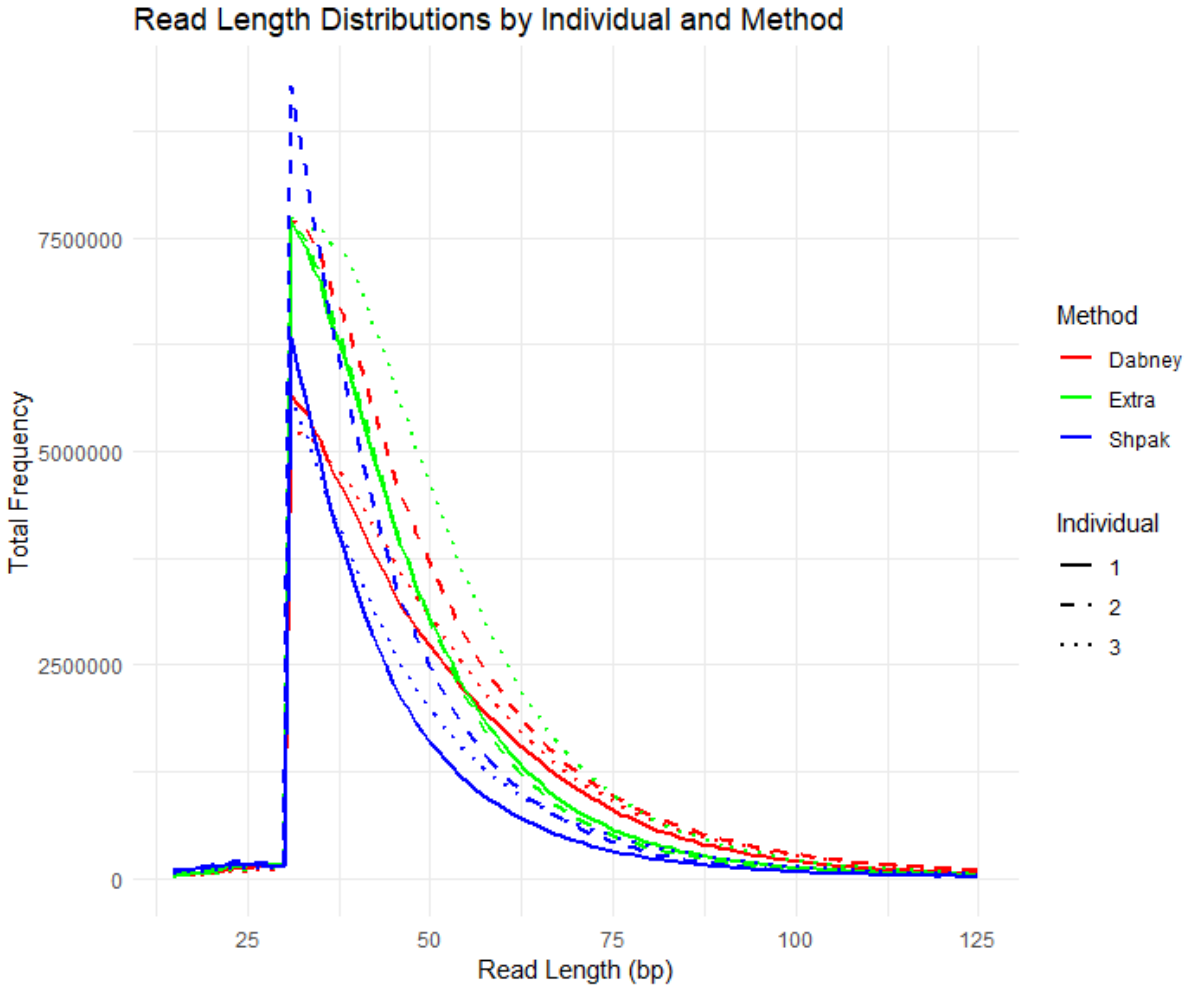

Figure 13: The sequence length distribution up to 125 bp (without binning) of the reads that were received from the deep sequencing of the three *B. germanica* individuals after trimming (Fig. 12). The x-axis was cut off after 125 bp due to better easier visibility and would else go up to 269 bp with very low frequencies. Partitions of the lysates (600 µL) were purified with either the Shpak or the Dabney protocol, which have different concentrations of carrier RNA during the binding of the DNA to the silica column. Additionally, the remaining lysates (130 µL for individual 1, 85 µL for individual 2 and 175 µL for individual 3) were purified to maximize the final output, leading to method "Extra". The samples that were purified with the Shpak protocol and a higher concentration of carrier RNA during the binding step led to a peak at a shorter sequence length. This suggests that carrier RNA may increase the retrieval of short DNA fragments which are otherwise lost during the lab procedures up to sequencing.

Using the mixed linear models to compare the two datasets obtained with the Shpak protocol – Shpak and "Extra" – but which are derived from different input volumes of lysate, we can once again see some differences. Having used the "Extra" method increased the read length by on average 1.86 bp ($P = 0.006996$) and the duplication rate by 4.16% ($P < 0.001$). Given that there will automatically be less DNA in the lower volumes of lysate, the decrease in complexity that is indicated by this higher duplication rate had to be expected. Yet surprisingly, the lower input volumes also increased the total yield of polished data by 53.46% ($P < 0.001$), which makes it the

method that retrieved the most data. This may suggest that the lysate input volume of 600 µL for the Shpak and Dabney protocols was too high. Considering that the measured DNA concentrations were the highest we ever measured for historical specimens, it might at first appear plausible that the columns have been overloaded. However, this can hardly be the case, as the MinElute Spin Columns are reported to hold up to 5 µg of DNA (Product specifications by QIAGEN), while the highest measured yield (purified from individual 2 with the Shpak protocol) was 0.094 µg. The only other notable difference between the Shpak and "Extra" lysates that occurred during the purification step – for all other steps the samples found themselves in mixed batches – was a minor alteration of the concentration of cRNA. According to the protocol, each 200 µL of lysate was supposed to receive 1 µg of cRNA before the binding to the column and this was precisely executed for the volumes of 600 µL by adding 3 µg. Yet, 1 µg still had to be added to the "Extra" volumes below 200 µL, as volumes below 1 µL could not have been pipetted with precision. Therefore, it can be speculated that the observed difference in yield has once again been caused by a difference in the concentration of cRNA. To gain certainty in this respect, it will be necessary to test different concentrations of cRNA with the otherwise same protocol and follow through with these samples all the way to deep sequencing. Plans to conduct such an experiment on a timeline of historical *B. germanica* individuals have been made but this could not have been included in this thesis.

Discussion

The effectiveness of the described method

For the three specimens of *B. germanica*, the developed method for non-destructive DNA extraction has been successful in achieving the two main goals of **(1)** ensuring the specimen´s physical integrity and **(2)** producing a satisfactory amount of sequence data on a whole genome level. The first goal has also been achieved in a wide variety of other taxa while the method´s exact limitations regarding the second goal cannot yet be conclusively determined. Doing so will require the deep sequencing of more specimens. However, all currently available evidence allows for optimism regarding the successful sequencing of historic drosophilids. This has already been achieved by Shpak et al. (2023) with 200-year-old *D. melanogaster* specimens and the results from a first trial sequencing run suggest that we have been successful with up to 110-year-old *D. simulans*.

This confidence is also strengthened by the finding that the addition of cRNA seems to have a positive effect on the yield by increasing the silica column's ability to retain short fragments of target DNA. Unfortunately, this effect could not be conclusively determined within the scope of this thesis because cRNA can be detected with our chosen method of DNA concentration measurement. However, the difference in read length distributions between the sequence data that comes from samples treated with the two tested extraction protocols (Fig. 13), indicates that cRNA has a positive effect. The moderate yet significant reduction in duplication rate that is shown to be achieved with the use of the Shpak protocol can be interpreted as an increase in complexity. This suggests that a larger proportion of the genome may be obtained using this method, further supporting the hypothesis that the use of cRNA is beneficial. Yet, the direct comparison between data obtained from Shpak and Dabney also shows that using the first relatively decreases the total amount of post-polished data by more than a third, which is an obvious downside and indicates a trade-off between complexity and sequencing cost. Still, another comparison between the two datasets that were obtained with the Shpak protocol – the ones where DNA was purified from the regular 600 µL partitions and the "Extra" ones that were purified from the remaining lysates – indicates that final data yield may simply be improved by using more cRNA. This could potentially remove the apparent cost of using cRNA and/or the Shpak protocol but the extent to which this is the case will have to be tested further in the future.

If these interpretations of our results are correct and the effects are reproducible, then this simple addition could be considered a significant improvement on Shpak et al. (2023) and cRNA may find wider application in the field of aDNA. For that specific use-case, this thesis is the first work to

report testing the efficacy of cRNA. Previous work on cRNA has reported mixed results. Further, we are the first to report that it may specifically help with the retrieval of DNA with shorter fragment sizes.

Morling (2009) reports that the addition of cRNA in QIAGEN DNA kits increased the yields by 24% on average. Positive effects have been achieved in various use cases, such as in, e.g., increasing the efficiency of robotic extractions for forensic purposes (Kishore et al. 2006), in miniaturized microfluidic applications (Shaw et al. 2009) and for embryo spent culture medium for the purpose of bioanalytical techniques in in-vitro fertilization (Sundari et al. 2021). Other studies report mixed results. El-Shorbagy et al. (2022) extracted DNA from bloodstains from two different kinds of fabric and they report that whether cRNA had a positive effect depended on the method of extraction as well as on sample condition. Reddy et al. (2012) found that it helped with the retrieval of DNA from fresh fecal matter of carnivores. Yet, this was not always the case for samples that have been stored, and they report that whether cRNA had a significant positive effect depends on storage protocol and duration. Further, Higgins et al. (2014) did not find any significant effect of cRNA on the yield of DNA extracted from “non-historic” teeth from human donors.

More testing is required to conclusively confirm that the addition of cRNA is advantageous for aDNA applications. In case our results with *B. germanica* can be reproduced with e.g., human bone, this may allow for the re-sequencing of material for which extraction has previously failed.

Considerations on the application

The used lysis buffer does not seem to cause significant visible damage to the specimens. Instead, physical damage in the form of the breaking-off of heads or appendages is directly associated with the handling of the specimens before and during the procedure. The small and large holding structures have been developed to mitigate this risk and help avoid physical damage as much as possible, but they must still be paired with maximal carefulness to be effective. For drosophilids, their implementation to avoid physical damage has been shown to be straightforward and this has succeeded for the vast majority of the many tested specimens. Working with larger taxa has overall been more challenging as the tested ones have observed very different levels of fragility. Yet, the finalized holding structures for “larger” specimens also allow us to perform a safe treatment, even though they are more difficult in handling. Finally, the best results in after-treatment appearance are obtained when the specimens are also carefully “re-preparated” shortly after the final wash.

Unfortunately, we lack understanding of how the treatment affects the inside of the specimens. However, for old material all informative inner structures will already have been long lost to the natural drying process. This makes them unusable for morphologists and at least in this way the effects the described treatment has on the inside of museum specimens are not important. Instead, practically all uses for historical specimens have been restricted to either classical morphology on the outside of the specimens, or more recently to genetic analyses. The latter could of course all be done on the same whole-genome-scale dataset, so long as quality and coverage of the sequencing output are sufficient. Yet, considering that the yields from fresh specimens were higher when lysis time was increased, it also seems likely that there should still be some DNA left after the usual 2 hours of lysis. This DNA could potentially still be accessible with a second extraction procedure.

However, it cannot be ruled out that the described treatment may prevent the use or bias the results of other methods in the future. Natural history collections represent a vast source of potential data, much of which was not intended at the time when they were established (Speer et al. 2022) and sometimes their uses can be quite surprising like Trumble et al. (2018) who analyzed the cortisol levels of whale earplugs to learn about stress responses to historic whaling. One use for natural history collections that may become relevant in the future is the sequencing of RNA. Only recently, Smith et al. (2019) have shown that it is possible for RNA to be preserved long-term by sequencing tissue-specific RNA from a 14,300-year-old permafrost-preserved canid. The possibility of sequencing ancient RNA of museum specimens has also been opened (Speer et al. 2022) and Keene & Stenglein (2024) extracted RNA from historical *D. melanogaster* specimens to detect retroviruses. For such an application, it seems likely that the lysis treatment described in this thesis would prohibit or bias the results.

Also, we do not know if the treatment may have negative long-term effects on the physical specimen. Lysing cells from connective tissues and muscles may well impair a specimen´s long term stability and make it more fragile. Considering that natural history collections are aiming for indefinite conservation, this should not be disregarded. The described method should thus not be employed on particularly valuable specimens, such as types. Gilbert et al. (2007) – who also performed DNA extractions by submerging the whole specimen in a lysis buffer – argue that important specimens should rather be leg-clipped to allow for future extraction. While this may no longer be an issue for DNA extractions if the whole genome has already been obtained from the first extraction, it might be an option for valuable specimens of larger taxa. This could not only guarantee the physical integrity of the voucher but also preserve the possibility of eventually performing molecular or chemical analyses that may become available in the future.

In the most optimal case, the described method should be applied to specimens when other syntopically collected ones exist and remain untreated. Nonetheless, considering the low risk that the described method poses to the specimen and how valuable historical genomes can be for evolutionary research – e.g., for reconstructing a species´ history of TE invasions (Pianezza et al. 2024b) and conservationist efforts (Jensen et al. 2022) – we propose that it can and should be used to extract the DNA of many specimens. Ultimately, curators and researchers will have to weigh the value of the sequenced genome – which is influenced by scientific, conservationist or medical relevance – against the potential risk and the value of maintaining the untreated specimen. The latter is mostly determined by its "rarity", which refers to both how common a species is found in nature and collections, as well as how many specimens there are from various regions and timeframes. When deciding whether to extract DNA from a specimen with the described method, one may keep in mind that any use of specimens, including for classical morphological work, poses some risk to it. Yet, a collection´s value comes precisely from it being able to be used and research that stems from this can help in advocating for the continued funding and maintenance of collections, which is particularly important in a time of rapid species loss (Shaffer et al. 1998).

Perspective and limitations on species

All current evidence suggests that the discussed protocol should also be applicable to other taxa. In particular, the described holding structure solutions should work for a large diversity of insect species. Naturally, this includes many species that we think could be interesting to investigate for their recent genomic history, especially in regards to potential TE invasions. As discussed previously in the introduction, it is not known what exactly triggers TE invasions or by what mechanism(s) they take place. A large comparative study by Muller et al (2024) across four phyla (annelids, arthropods, mollusks and chordates) found that the likelihood of HTTs between two species is increased by phylogenetic relatedness and similarity of habitat types, while Peccoud et al. (2017) report a correlation with geographic origin in insects overall. In the case of *D. melanogaster*, Scarpa et al (2024) point to a large habitat expansion with subsequent secondary contact with closely related species such as *D. willistoni*, as well as to a general environmental shift caused by close human commensalism as potential explanations.

However, there are a few more factors that may plausibly increase the chance and thereby historic rate of HTTs. For once, we hypothesize that close physical contact with other animals could be an important factor. Thus, recent invasions may particularly be found in species with complex life histories that are engaged in an expansive net of interspecific biological interaction. Additionally,

it is plausible that a large number of reproducing individuals – this excludes eusocial (Hymenopteran) species such as *Apis melifera*, *Vespula germanica* and all ants – also directly increases the chance for horizontal transfer. Further, as genome size diversity in insects is largely determined by TE content (Maumus et al. 2015), one may expect to find high rates of HTT in species that have large genomes with many different families of TEs. One interesting candidate would thus be the migratory locust *Locusta migratoria*, as it possesses the to date largest genome of all insects, which is directly tied to TE proliferation (Wang et al. 2014).

The potential of the described method is of course directly limited by the access to and existence of historical pinned material. For specimens of groups that are not commonly pinned, the lysis treatment with the holding structures is not applicable, while the rest of the described extraction and purification procedure (everything onwards from DNA purification) may still be followed. Yet, even for pinned specimens there are some limitations regarding the size and forms of species and ones that possess features that are likely to be destroyed by the lysis treatment are not suitable for DNA extraction with the described method. One example for such a group are lepidopterans because of their scales (Patzoldi et al. 2020), although they would be the hotspot of HTT in insects (Muller et al. 2024). Moreover, there are many species that are too small to be able to receive a satisfactory yield of DNA using this method. For many of these species they are practically too small to be pinned or there are no pinned specimens available due to historical reasons. Instead, such species are often glued to small white paper strips. This excludes the whole group of aphids which would otherwise be very interesting to investigate due to their role as plant disease vectors (Nault 1997). Although we do not yet have replicated Shpak et al (2023) by deep sequencing historical drosohilids that are up to 200 years old, we are confident that we should be able to. The size of *D. melanogaster* can thus be assumed to be the fairly confident lower limit. However, there are generally not that many species smaller than *D. melanogaster* that are pinned in historical collections.

For bigger specimen DNA extraction yields are no concern. Rather, the holding structures are the limitation. The size of the container of the washes and the volumes of the washes could very easily be expanded and the specimens´ safety could be ensured by modifying the holding structure. Yet, if a Combitip is used as a container for the lysis buffer, the distance from the pin to the furthest part of the specimens needs to be smaller than the Combitip´s radius of approximately 2 cm. However, it appears certain that another solution could eventually be found (e.g., by 3D printing) and would allow extracting from larger species, such as from orthopterans like *Locusta migratoria*.

References

- Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. [Version 0.11.9]
- Asghar, U., Malik, M.F., Anwar, F., Javed, A. & Raza, A. 2015. DNA Extraction from Insects by Using Different Techniques: A Review. Advances in Entomology 3:132-138. doi: 10.4236/ae.2015.34016.
- Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67(1):1–48. doi: 10.18637/jss.v067.i01.
- Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., Mager, D.L. & Feschotte, C. 2018. Ten things you should know about transposable elements. Genome Biology 19:199. doi: 10.1186/s13059-018-1577-z.
- Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R. & Hannon, G.J. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. 2007. Cell 128:1089–1103. doi: 10.1016/j.cell.2007.01.043.
- Brown, R.B. & Audet, J. 2008. Current techniques for single-cell lysis. Journal of the Royal Society Interface 5:131–138. doi: 10.1098/rsif.2008.0009.focus.
- Chen, S. 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta 2:e107. doi: 10.1002/imt2.107.
- Chen, S., Zhou, Y., Chen, Y. & Gu, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34(17):i884–i890. doi: 10.1093/bioinformatics/bty560.
- Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L. & Meyer, M. 2013. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proceedings of the National Academy of Sciences 110 (39):15758-15763. doi: 10.1073/pnas.1314445110.
- Daborn, P.J., Yen, J.L., Bogwitz, M.R., Le Goff, G., Feil, E., Jeffers, S., Tijet, N., Perry, T., Heckel, D., Batterham, P., Feyereisen, R., Wilson, T.G. & ffrench-Constant, R.H. 2002. A Single P450 Allele Associated with Insecticide Resistance in *Drosophila*. Science 297(5590):2253-2256. doi: 10.1126/science.1074170.
- El-Shorbagy, H.M., El-Liethy, S.S., Moussa, M.K. & Elghor, A.A. 2022. Carrier RNA is a key factor affecting fully integrated short tandem repeats profiling in challenging forensic samples models. The Journal of Basic and Applied Zoology 83:24. doi: 10.1186/s41936-022-00286-7.

- Ewels, P., Magnusson, M., Lundin, S. & Käller, M. 2016. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32(19):3047-8. doi: 10.1093/bioinformatics/btw354. [Version 1.27.1]
- Fong, J.J., Blom, M.P.K., Aowphol, A., McGuire, J.A., Sutcharit, C. & Soltis, P.S. 2023. Editorial: Recent advances in museomics: revolutionizing biodiversity research. Frontiers in Ecology and Evolution 11:1188172. doi: 10.3389/fevo.2023.1188172.
- Gilbert, M.T.P., Moore, W., Melchior, L. & Worobey, M. 2007. DNA Extraction from Dry Museum Beetles without Conferring External Morphological Damage. PLoS ONE 2(3):e272. doi: 10.1371/journal.pone.00002722007.
- Glocke, I. & Meyer, M. 2017. Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. Genome Research 27(7):1230–1237. doi: 10.1101/gr.219675.116.
- Higgins, D., Kaidonis, J., Townsend, G. & Austin, J.J. 2014. Evaluation of carrier RNA and low volume demineralization for recovery of nuclear DNA from human teeth. Forensic Science Medicine and Pathology 10:56–61. doi: 10.1007/s12024-013-9519-2.
- Jensen, E.L., Díez-del-Molino, D., Gilbert, M.T., Bertola, L.D., Borges, F., Cubric-Curik, V., de Navascués, M., Frandsen, P., Heuertz, M., Hvilsom, C., Jiménez-Mena, B., Miettinen, A., Moest, M., Pečnerová, P., Barnes, I. & Vernesi, C. 2022. Ancient and historical DNA in conservation policy. Trends in Ecology & Evolution 37(5):420-428. doi: 10.1016/j.tree.2021.12.010.
- Kapp, J.D., Green, R.E. & Shapiro, B. 2021. A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. Journal of Heredity 112(3):241–249. doi: 10.1093/jhered/esab012.
- Kassambara, A. 2020. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr.
- Keene, A.H. & Stenglein, M.D. 2024. Viral metagenomics of 100-year-old museum specimens highlights the long-term stability of RNA. bioRxiv preprint. doi: 10.1101/2024.10.03.616531.
- Kidwell, M.G. 1982. Evolution of hybrid dysgenesis determinants in *Drosophila melanogaster*. Proceedings of the National Academy of Sciences 80:1655-1659. doi: 10.1073/pnas.80.6.1655.
- Kishore, R., Hardy, W.R., Anderson, V.J., Sanchez, N.A. & Buoncristiani, M.R. 2006. Optimization of DNA Extraction from Low-Yield and Degraded Samples Using the BioRobot® EZ1 and BioRobot® M48. Journal of Forensic Sciences 51(5):1055-1061. doi: 10.1111/j.1556-4029.2006.00204.x.

- Kuznetsova, A., Brockhoff, P.B. & Christensen, R.H.B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software 82(13):1–26. doi: 10.18637/jss.v082.i13.
- Lindahl, T. 1993. Instability and decay of the primary structure of DNA. Nature 362:709-715.
- Kircher, M., Sawyer, S. & Meyer, M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Research 40(1):e3 doi: 10.1093/nar/gkr771.
- Mandrioli, M. 2008. Insect collections and DNA analyses: How to manage collections? Museum Management and Curatorship 23:193–199. doi: 10.1080/09647770802012375.
- Maumus, F., Fiston-Lavier, A-S. & Quesneville, H. 2015. Impact of transposable elements on insect genomes and biology. Current Opinion in Insect Science 7:30-36. doi: 10.1016/j.cois.2015.01.001.
- McClintock, B. 1950. The origin and behavior of mutable loci in maize. Proceedings of the National Academy of Sciences 36(6):344–355. doi: 10.1073/pnas.36.6.344.
- Morling, N. 2009. PCR in forensic genetics. Biochemical Society Transactions 37(2):438–440. doi: 10.1042/BST0370438.
- Muller, H., Savisaar, R., Peccoud, J., Charlat, S. & Gilbert, C. 2024. Phylogenetic relatedness rather than aquatic habitat fosters horizontal transfer of transposable elements in animals. bioRxiv preprint doi: 10.1101/2024.12.18.629015.
- Nault, N.R. 1997. Arthropod Transmission of Plant Viruses: A New Synthesis. Annals of the Entomological Society of America 90(5):521-541. doi: 10.1093/aesa/90.5.521.
- New England Biotech. 2020 Genomic DNA Extraction from Insects. Link: https://www.neb.com/en/protocols/2020/02/04/genomic-dna-purification-from-insects [16/9/2024]
- Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P.W., Ávila-Arcos, M.C., Fu, Q., Krause, J., Willerslev, E., Stone, A.C. & Warinner, C. 2021. Ancient DNA analysis. Nature Reviews Methods Primers 1:14. doi: 10.1038/s43586-020-00011-0.
- Patzoldi, F., Zilli, A. & Hundsdörfer, A.K. 2020. Advantages of an easy-to-use DNA extraction method for minimal-destructive analysis of collection specimens. PLoS ONE 15(7):e0235222. doi: 10.1371/journal.pone.0235222.
- Peccoud, J., Loiseau, V., Cordaux, R. & Gilbert, C. 2017. Massive horizontal transfer of transposable elements in insects. Proceedings of the National Academy of Sciences 114(18):4721-4726. doi: 10.1073/pnas.1621178114.

- Periquet, G., Hamelin, M.H., Bigot, Y., & Lepissier, A. 1989. Geographical and historical patterns of distribution of hobo elements in *Drosophila melanogaster* populations. Journal of Evolutionary Biology 535(3):223–229. doi: 10.1046/j.1420-9101.1989.2030223.x.
- Phillips, A.J. & Simon, C. 1995. Simple, Efficient, and Nondestructive DNA Extraction Protocol for Arthropods. Annals of the Entomological Society of America 88(3):281-283. doi: 10.1093/aesa/88.3.281.
- Pianezza, R., Scarpa, A., Narayanan, P., Signor, S. & Kofler, R. 2024 (a). Spoink, a LTR retrotransposon, invaded *D. melanogaster* populations in the 1990s. PLOS Genetics 20(3):e1011201. doi: 10.1371/journal.pgen.1011201.
- Pianezza, R., Scarpa, A., Haider, A., Signor, S. & Kofler, R. 2024 (b). Unveiling the complete invasion history of D. melanogaster: three horizontal transfers of transposable elements in the last 30 years. bioRxiv preprint. doi: 10.1101/2024.04.25.59109.
- Pianezza, R., Haider, A. & Kofler, R. 2024 (c). GenomeDelta: detecting recent transposable element invasions without repeat library. bioRxiv preprint. doi: 10.1101/2024.06.28.601149.
- QIAGEN 2003. Protocol: Isolation of Genomic DNA from Tissues. QIAamp DNA Micro Handbook 08/2003.
- R Core Team 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. [Version 4.2.2 (2022 10-31 ucrt)]
- Reddy, P.A., Bhavanishankar, M., Bhagavatula, J., Harika, K., Mahla, R.S. & Shivalji, S. 2012. Improved Methods of Carnivore Faecal Sample Preservation, DNA Extraction and Quantification for Accurate Genotyping of Wild Tigers. PLoS ONE 7(10):e46732. doi: 10.1371/journal.pone.0046732.
- Santos, D., Ribeiro, G.C., Cabral, A.D. & Sperança, M.A. 2018. A non-destructive enzymatic method to extract DNA from arthropod specimens: Implications for morphological and molecular studies. PLoS ONE 13(2):e0192200. doi: 10.1371/journal.pone.0192200.
- Scarpa, A., Pianezza, R., Wierzbicki F. & Kofler, R. 2024. Genomes of historical specimens reveal multiple invasions of LTR retrotransposons in *Drosophila melanogaster* during the 19th century. Proceedings of the National Academy of Sciences 121(15):e2313866121. doi: 10.1073/pnas.2313866121.

- Schaak, S., Gilbert, C. & Feschotte, C. 2010. Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends in Ecology and Evolution 25:537-546. doi: 10.1016/j.tree.2010.06.001.
- Schwarz, F., Wierzbicki, F., Senti, K.-A., & Kofler, R. 2020. Tirant stealthily invaded natural *Drosophila melanogaster* populations during the last century. Molecular Biology and Evolution 38(5). doi: 10.1093/molbev/msaa308.
- Shaffer, H.B., Fisher, R.N. & Davidson, C. 1998. The role of natural history collections in documenting species declines. TREE 13(1):27-30. doi: 10.1016/S0169-5347(97)01177-4.
- Shaw, K.J., Thain, L., Docker, P.T., Dyer, C.E., Greeman, J., Greenway, G.M. & Haswell, S.J. 2009. The use of carrier RNA to enhance DNA extraction from microfluidic-based silica monoliths. Analytica Chimica Acta 652:231–233. doi: 10.1016/j.aca.2009.03.038.
- Shen, W., Le, S., Li, Y. & Hu, F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE 11(10):e0163962. doi: 10.1371/journal.pone.0163962.
- Shpak, M., Ghanavi, H.R., Lange, J.D., Pool, J.E. & Stensmyr, M.C. 2023. Genomes from historical Drosophila melanogaster specimens illuminate adaptive and demographic changes across more than 200 years of evolution. PLoS Biology 21(10):e3002333. doi: 10.1371/journal.pbio.3002333.
- Slatkin, M. & Racimo, F. 2016. Ancient DNA and human history. Proceedings of the National Academy of Sciences 113(23):6380–6387. doi: 10.1073/pnas.1524306113.
- Smith, O., Dunshea, G., Sinding, M-H.S., Fedorov, S., Germonpre, M., Bocherens, H. & Gilbert, M.T.P. 2019. Ancient RNA from Late Pleistocene permafrost and historical canids shows tissue-specific transcriptome survival. PLoS Biol 17(7):e3000166. doi: 10.1371/journal.pbio.3000166.
- Speer, K.A., Hawkins, M.T.R., Flores, M.F.C., McGowen, M.R., Fleischer, R.C., Maldonado, J.E., Campana, M.G. & Muletz-Wolz, C.R. 2022. A comparative study of RNA yields from museum specimens, including an optimized protocol for extracting RNA from formalin-fixed specimens. Frontiers in Ecology and Evolution 10:953131. doi: 10.3389/fevo.2022.953131.
- Strunov, A., Kirchner, S., Schindelar, J., Kruckenhauser, L., Haring, E. & Kapun, M. 2023. Historic Museum Samples Provide Evidence for a Recent Replacement of *Wolbachia* Types in European *Drosophila melanogaster*. Molecular Biology and Evolution 40(12):msad258. doi: 10.1093/molbev/msad258.
- Sundari, A.M., Iffanolida, P.A., Mutia, K., Muna, N., Mansyur, E., Hestiantoro, A., Wiweko, B., Fauzi, M. & Bowolaksono, A. 2021. Carrier RNA (cRNA) Enhances dsDNA Recovery

Extracted from Small Volume Spent Embryo Culture Medium. International Journal on Advanced Science Engineering and Information Technology 11(3):1203-1207. doi: 10.18517/ijaseit.11.3.13591.

- Tang, Q., Bourguignon, T., Willenmse, L., De Coninck, E. & Evans, T. 2019. Global spread of the German cockroach, Blattella germanica. Biological Invasions 21:693–707. doi:10.1007/s10530-018-1865-2.
- Tang, Q., Vargo, E., Ahmad, I., Jiang, H., Varadínová, Z., Dovih, P., Kim, D., Bourguignon, T., Booth, W., Schal, C., Mukha, D., Rheindt, F. & Evans, T. 2024. Solving the 250-year-old mystery of the origin and global spread of the German cockroach, *Blattella germanica*. Proceedings of the National Academy of Sciences of the United States of America 121(22):e2401185121. doi: 10.1073/pnas.2401185121.
- Trumble, S.J., Norman, S.A., Crain, D.D., Mansouri, F., Winfield, Z.C., Sabin, R., Potter, C.W., Gabriele, C.M. & Usenko, S. 2018. Baleen whale cortisol levels reveal a physiological response to 20th century whaling. Nature Communications 9:4587. doi: 10.1038/s41467-018-07044-w.
- van´t Hof, A.E., Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., Hall, N., Darby, A.C. & Saccheri, I.J. 2016. The industrial melanism mutation in British peppered moths is a transposable element. Nature 534:102-105. doi: 10.1038/nature17951.
- Wang, X., Fang, X., Yang, P., Jiang, X., Jiang, F., Zhao, D., Li, B., Cui, F., Wei, J., Ma, C., Wang, Y., He, J., Luo, Y., Wang, Z., Guo, X., Guo, W., Wang, X., Zhang, Y., Yang, M., Hao, S., Chen, B., Ma, Z., Yu, D., Xiong, Z., Zhu, Y., Fan, D., Han, L., Wang, B., Chen, Y., Wang, J., Yang, L., Zhao, W., Feng, Y., Chen, G., Lian, J., Li, Q., Huang, Z., Yao, X., Lv, N., Zhang, G., Li, Y., Wang, J., Wang, J., Zhu, B. & Kang, L. 2014. The locust genome provides insight into swarm formation and long-distance flight. Nature Communications 5:2957. doi: 10.1038/ncomms3957
- Wells, J.N. & Feschotte, C. 2020. A Field Guide to Eukaryotic Transposable Elements. Annual Review of Genetics 54:539–561. doi: 10.1146/annurev-genet-040620-022145.
- Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. 2023. dplyr: A Grammar of Data Manipulation. R package version 1.1.4, Link: https://github.com/tidyverse/dplyr, https://dplyr.tidyverse.org.

Appendix

Alternative DNA purification of the lysate with the Dabney protocol

The following protocol is modified from an internal lab protocol of the Pinhasi group.

Reagents:

- Binding buffer – recipe for 50 mL (make sure the ingredients are added in this order; should be stored up to one month):
    - Guanidine hydrochloride (MW 95.53): 23.9 g
    - Isopropanol: 20 mL
    - Sodium Acetate pH 5.2 (3M): 1.5 mL
    - Tween-20 (10%): 250 µL
    - Fill up to 50 mL with H_2O
- PE buffer

1. If you have had a **"larger" specimen**, determine the volume of lysate (**v**) that you have in µL. If you have had a **"smaller" specimen** it will be **v = 200**.
2. Transfer the lysate into a High Pure Extender Assembly.
3. Add a total of **13 v µL of binding buffer** and **v/200 µl dissolved carrier RNA** to the lysate. Centrifuge the High Pure Extender Assembly for 4 min at 1,500 rcf. If needed, rotate 90º and centrifuge again for 2 min at the same speed.
4. Remove the extension reservoir from the High Pure Extender Assembly and place into a fresh collection tube. Pull the tab to remove the extender and break off, leaving only the spin column. Label the lid of the spin column tubes.
5. Dry spin for 1 min at 6,000 rpm to remove any remaining binding buffer, then place the spin column into a fresh collection tube.
6. Add **650 µL of PE washing buffer** to the spin column and centrifuge again for 1 min at 6,000 rpm, discarding the flow-through. Repeat this washing step.
7. Dry spin the spin column at full speed (14,000 rpm) for 1 min to remove any washing buffer. This step is necessary since ethanol carryover into the eluate may interfere with downstream applications.
8. Place the column in a (labelled) clean 1.5 mL microcentrifuge tube and discard the collection tube containing the flow-through.

9. **First step of elution:** Carefully open the lid of the column and apply **20 µl Buffer AE** to the center of the membrane. The volume and the elution buffer can be varied.
10. Close the lid and incubate at RT for 15 min. Centrifuge at full speed (14,000 rpm) for 1 min. You now have the **primary eluate**.
11. **Second step of elution:** If you want to have the secondary eluate in a separate tube, place the column in a clean 1.5 mL microcentrifuge tube. Carefully open the lid of the column and apply **20 µl Buffer AE** to the center of the membrane. The volume and the elution buffer can be varied.
12. Close the lid and incubate at RT for 15 min. Centrifuge at full speed (14,000 rpm) for 1 min.
13. Discard the column. You now have the **secondary eluate**. It will contain some more DNA which may either be pooled with the primary one or be kept as a backup.