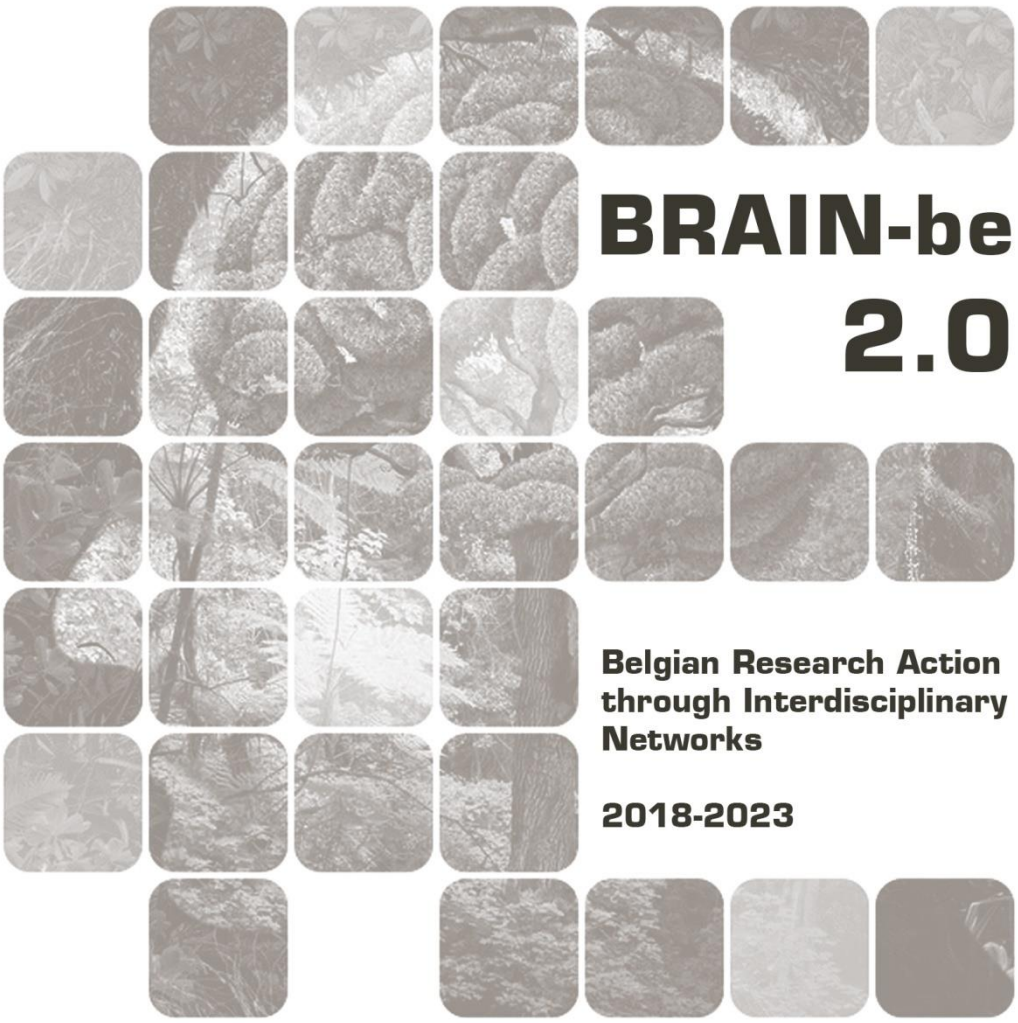

InsectMOoD

Insect Museum Open -omic Database

Massimiliano Virgilio, Royal Museum for Central Africa (RMCA)

Carl Vangestel, Royal Belgian Institute of Natural Sciences (RBINS)

Pillar 2: Heritage science

belspo .be

NETWORK PROJECT

InsectMOoD

Insect Museum Open -omic Database

Contract - B2/202/P2/InsectMOoD

FINAL REPORT

PROMOTORS:

Massimiliano Virgilio (RMCA)

Carl Vangestel (RBINS)

AUTHORS:

Lore Esselens (RMCA)

Gontran Sonet (RBINS)

Carl Vangestel (RBINS)

Massimiliano Virgilio (RMCA)

Published in 2023 by the Belgian Science Policy Office

WTCIII

Simon Bolivarlaan 30 bus 7

Boulevard Simon Bolivar 30 bte 7

B-1000 Brussels

Belgium

Tel: +32 (0)2 238 34 11

http://www.belspo.be
http://www.belspo.be/brain-be

Contact person: Georges JAMART

Tel: +32 (0)2 238 36 90

InsectMOoD. Final Report. Brussels: Belgian Science Policy Office 2023 – 52 p. (BRAIN-be 2.0 - (Belgian Research Action through Interdisciplinary Networks))

TABLE OF CONTENTS

ABSTRACT

Context

Museum vouchers from biological collections are of particular importance for scientific research on taxonomy, systematics and biogeography and provide tools to tackle a wide range of scientific questions in disciplines such as ecology, evolution and conservation. The rapid technological advances over the past few years have led to a substantial reduction in costs, so that routine high throughput sequencing (HTS) of collection vouchers, including their whole genome sequencing (WGS), represents an exciting perspective for the valorisation of museum collections. However, museum specimens are often not directly suitable for genetic/genomic analyses due to low-quality DNA.

Objectives

InsectMOoD aimed at (a) providing a feasibility study for the large-scale WGS of Diptera from museum collections, (b) explore suitable approaches to create open-access, economically affordable and ready-to-use databases and repositories of genomic resources and (c) develop a decision map for the routine archiving of *genomic vouchers* as a complement to the routine archiving of morphological and digital vouchers at the Royal Museum for Central Africa. In this context, a main focus was given to hoverflies (Diptera, *Syrphidae*) and "true" fruit flies (Diptera, *Tephritidae*), two taxon groups for which RMCA has considerable taxonomic expertise and ongoing collection-based research.

Methods and Results

This project provided encouraging indications about the large-scale collection of genomic data from insect museum vouchers and resulted in the production of a consistent amount of "genomic vouchers" (> 1,300). These latter were represented by whole genomes of collection vouchers and by their metadata (including, *inter alia*, information on protocols used for genomic library preparation and high throughput sequencing). The genomic vouchers produced were linked to morphological and digital vouchers as well as to the associated DNA collection vouchers. InsectMOoD indicates the suitability of a two-step approach, which consists in the use of (a) commercial kits and standard genomic library preparation protocols for cost- and time-effective genotyping of subsets of Museum vouchers and (b) more specialised protocols (e.g. including aDNA methodologies) to be used exclusively for the more problematic subsets of specimens which did not yield satisfactory results with routine methodologies.

Conclusions

We believe that InsectMOoD provided a remarkable added value to the insect collections of RMCA . This approach allowed delivering a large bulk of easily accessible genetic information available in the framework of ongoing and future research on *Tephritidae* and *Syrphidae*. The optimization of experimental protocols and the collection of the genomic data, coordinated by the Joint Experimental Molecular Unit (JEMU) of RMCA and RBINS, generated guidelines of general interest for the WGS genotyping of material from the biological collections of RMCA and RBINS and further strengthened the expertise of the JEMU in Museomics. The multi-layer collection system advocated by this project will allow upgrading the RMCA standards on the collection of genomic data from museum vouchers.

(Project executive summary and French and Dutch versions of this abstract provided in Annex 1)

Keywords

Museum collections, Genomic Voucher, Collection management, High Throughput Sequencing, Whole Genome Sequencing

1. INTRODUCTION

Vouchers from natural history collections represents a vast repository of biodiversity. Advances in laboratory and sequencing technologies have made these specimens increasingly accessible for genomic analyses, offering a window into the genetic past of species and often permitting access to information that can no longer be sampled in the wild. Due to their age, preparation and storage conditions, DNA retrieved from museum and herbarium specimens is often poor in yield, heavily fragmented and biochemically modified. This not only poses methodological challenges in recovering nucleotide sequences, but also makes such investigations susceptible to environmental and laboratory contamination (Ferrari *et al.*, 2023).

2. STATE OF THE ART AND OBJECTIVES

The continuous progresses in genomic technologies keeps on providing new tools for the genetic characterisation of historical samples in ways that were not imaginable until only a few years ago (Colella, Tigano and MacManes, 2020). In this respect, an increasing number of dedicated -omic protocols for sub-optimal or ancient DNA from Museum vouchers (museomics) have been developed with the specific objective of mining genomic data from Natural History Collections (Guschanski *et al.*, 2013; Knyshov, Gordon and Weirauch, 2019; Knyshov, Hoey-Chamberlain and Weirauch, 2019). Yet, even if many of the proposed methodologies allow recovering highly degraded genetic material from ancient specimens, they are often too articulated and time consuming and that they would not be economically sustainable for the large-scale genotyping of museum vouchers. Hence, we advocate the use of a pragmatic approach to the routine genotyping of suboptimal Museum vouchers as, very often, they represent a consistent part of the collection vouchers. So far, the costs directly related to genomic library preparation and sequencing represented one of the main limiting factors hampering the whole genome sequencing (WGS) of large number of vouchers and, until recently, the partial sequencing of genomes, e.g. via reduced representation genomic libraries (Ewart *et al.*, 2019) or mitochondrial genomics (Timmermans *et al.*, 2016), was considered, as the only suitable approach to build up relatively large genomic datasets. However, the rapid technological advances over the past few years, have now led to a substantial reduction in costs, so that the routine WGS of vouchers represents a new, exciting perspective for the valorisation of Museum collections (Crampton-Platt *et al.*, 2016; Malakasi *et al.*, 2019; Strijk *et al.*, 2020).

The biological collections of the Royal Museum for Central Africa (RMCA) represent valuable repositories of vouchers collected over the past 150 years in the framework of a wide range of scientific expeditions and research projects. These collections include an estimated amount of six million insect specimens potentially available for research on taxonomy and systematics, biodiversity conservation, insect pest control and pollination ecology. The tephritid and syrphid collections of RMCA include more than 100,000 samples and, as a consequence of the research activities of specialized taxonomists actively involved in national and international collaborations, are among the most intensively exploited collections of RMCA. The digitalization of the RMCA insect collections has been the topic of consecutive programs (and including DIGIT-03, DIGIT-04, 3DSPECTRAL, DiSSCo-FED) and allowed converting a large number of morphological vouchers into digital vouchers that can now be accessed by a larger public. However, with the possible exclusion of the efforts made to establish a collection of DNA extracts, considerably less effort has been put in valorising their impressive bulk of genomic resources, although ready-to-use genomic data could be of great interest in the context of fundamental or applied research.

In this project, we deal with the practical challenges associated to the recovery of genomic data from museum collections with the general objective of promoting the large-scale genotyping of Museum vouchers as a routine preventive or curative intervention to preserve and valorise the collection genetic resources. This approach would advance the production of “genomic vouchers” represented by whole genomes of collection vouchers and by their metadata (including, *inter alia*, information on protocols used for genomic library preparation and high throughput sequencing) linked to the

corresponding morphological and digital vouchers as well as to the associated DNA collection voucher in a multi-layer voucher collection system.

The specific objectives of this two year project were to (a) provide a test-case for the creation of genomic collections of Diptera at RMCA (in addition to the morphological and digital collections), (b) explore suitable approaches to create open-access, economically affordable and ready-to-use databases and repositories of genomic resources and (c) develop a decision map for the routine archiving of *genomic vouchers* as a complement to the routine archiving of morphological and digital vouchers. In this context, a main focus was given to hoverflies (Diptera, *Syrphidae*) and "true" fruit flies (Diptera, *Tephritidae*), two taxon groups for which RMCA has considerable taxonomic expertise and ongoing collection-based research.

3. METHODOLOGY AND SCIENTIFIC RESULTS

WP1: Analysis of relationships between DNA quality and quantity and WGS performance in Museum vouchers.

Task 1.1. Review of the available experimental data

Information about the latest wet lab pipelines for DNA extraction, genomic library preparation and high throughput sequencing of Museum vouchers was retrieved from the scientific literature (see lists of references in the end of this section and in the end of this report). Wet-lab pipelines and experimental protocols used over the last few years by the Joint Experimental Molecular Unit of RMCA and RBINS (see http://jemu.myspecies.info/projects) were preliminarily scrutinised, discussed, and considered with respect to their expected suitability for the routine genotyping of Museum vouchers.

Task 1.2. Selection of suitable methodological approaches for routine Museomics

Methods were preliminarily evaluated in terms of pipeline complexity and expected costs with priority given to less articulated and less expensive pipelines (as better suitable for the routine processing of large numbers vouchers). More specialized protocols were excluded from consideration as already considered in the framework of dedicated procedures for the WGS of degraded vouchers of particular relevance. Suitable wet lab pipelines for the routine DNA extraction, genomic library preparation and high throughput sequencing of Museum vouchers were scrutinized and selected, and their performances experimentally quantified and compared.

D 1.1.1: report on suitable methodological literature (M4)

The preliminary literature revision performed in the first four months of the project provided a first selection of papers which were deemed relevant to the InsectMOoD objectives:

- Besnard, G. *et al.* (2016) 'Valuing museum specimens: High-throughput DNA sequencing on historical collections of New Guinea crowned pigeons (Goura)', *Biological Journal of the Linnean Society*, 117(1), pp. 71–82. doi: 10.1111/bij.12494.
- Bi, K. *et al.* (2013) 'Unlocking the vault: Next-generation museum population genomics', *Molecular Ecology*, 22(24), pp. 6018–6032. doi: 10.1111/mec.12516.
- Blaimer, B. B. *et al.* (2016) 'Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens', *PLoS ONE*, 11(8), p. 161531. doi: 10.1371/journal.pone.0161531.
- Buenaventura, E. (2021) 'Museomics and phylogenomics with protein-encoding ultraconserved elements illuminate the evolution of life history and phallic morphology of flesh flies (Diptera: Sarcophagidae)', *BMC Ecology and Evolution*, 21(1). doi: 10.1186/s12862-021-01797-7.
- Burrell, A. S., Disotell, T. R. and Bergey, C. M. (2015) 'The use of museum specimens with high-throughput DNA sequencers', *Journal of Human Evolution*, 79, pp. 35–44. doi: 10.1016/j.jhevol.2014.10.015.

- Call, E. *et al.* (2021) 'Museomics: Phylogenomics of the Moth Family Epicopeiidae (Lepidoptera) Using Target Enrichment', *Insect Systematics and Diversity*, 5(2). doi: 10.1093/isd/ixaa021.
- Cridland, J. M. *et al.* (2018) 'Genome Sequencing of Museum Specimens Reveals Rapid Changes in the Genetic Composition of Honey Bees in California', *Genome Biology and Evolution*, 10(2), pp. 458–472. doi: 10.1093/gbe/evy007.
- Faircloth, B. C. *et al.* (2015) 'Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among hymenoptera', *Molecular Ecology Resources*, 15(3), pp. 489–501. doi: 10.1111/1755-0998.12328.
- Gauthier, J. *et al.* (2020) 'Museomics identifies genetic erosion in two butterfly species across the 20th century in Finland', in *Molecular Ecology Resources*. Blackwell Publishing Ltd, pp. 1191–1205. doi: 10.1111/1755-0998.13167.
- Gillett, C. P. D. T. *et al.* (2014) 'Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea)', *Molecular Biology and Evolution*, 31(8), pp. 2223–2237. doi: 10.1093/molbev/msu154.
- Guschanski, K. *et al.* (2013) 'Next-generation museomics disentangles one of the largest primate radiations', *Systematic Biology*, 62(4), pp. 539–554. doi: 10.1093/sysbio/syt018.
- Hung, C. M. *et al.* (2013) 'The De Novo Assembly of Mitochondrial Genomes of the Extinct Passenger Pigeon (Ectopistes migratorius) with Next Generation Sequencing', *PLoS ONE*, 8(2), p. 56301. doi: 10.1371/journal.pone.0056301.
- Mayer, C. *et al.* (2021) 'Adding leaves to the Lepidoptera tree: capturing hundreds of nuclear genes from old museum specimens', *Systematic Entomology*, 46(3), pp. 649–671. doi: 10.1111/syen.12481.
- Mikheyev, A. S. *et al.* (2017) 'Museum Genomics Confirms that the Lord Howe Island Stick Insect Survived Extinction', *Current Biology*, 27(20), pp. 3157-3161.e4. doi: 10.1016/j.cub.2017.08.058.
- Papanicolaou, A. *et al.* (2016) 'The whole genome sequence of the Mediterranean fruit fly, Ceratitis capitata (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species', *Genome Biology*, 17(1), pp. 1–31. doi: 10.1186/s13059-016-1049-2.
- Rohland, N., Siedel, H. and Hofreiter, M. (2010) 'A rapid column-based ancient DNA extraction method for increased sample throughput', *Molecular Ecology Resources*, 10(4), pp. 677–683. doi: 10.1111/j.1755-0998.2009.02824.x.
- Rowe, K. C. *et al.* (2011) 'Museum genomics: Low-cost and high-accuracy genetic data from historical specimens', *Molecular Ecology Resources*, 11(6), pp. 1082–1092. doi: 10.1111/j.1755-0998.2011.03052.x.
- Suchan, T. *et al.* (2016) 'Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens', *PLoS ONE*, 11(3), p. e0151651. doi: 10.1371/journal.pone.0151651.
- Timmermans, M. J. T. N. *et al.* (2016) 'Rapid assembly of taxonomically validated mitochondrial genomes from historical insect collections', *Biological Journal of the Linnean Society*, 117(1), pp. 83–95. doi: 10.1111/bij.12552.
- Tin, M. M. Y., Economo, E. P. and Mikheyev, A. S. (2014) 'Sequencing degraded DNA from non-destructively sampled museum specimens for RAD-tagging and low-coverage shotgun phylogenetics', *PLoS ONE*, 9(5), p. 96793. doi: 10.1371/journal.pone.0096793.
- Waku, D. *et al.* (2016) 'Evaluating the phylogenetic status of the extinct Japanese otter on the basis of mitochondrial genome analysis', *PLoS ONE*, 11(3), p. 149341. doi: 10.1371/journal.pone.0149341.

D 1.1.2: report on suitable JEMU datasets for validation of results (M4)

An additional dataset from the JEMU project KEARAD, including DNA extracts from cichlid fishes dating from 1984 to 2019 from the collections of RMCA, was initially considered of possible interest to the project objectives (see project report 2022). In fact, these data seemed to suggest a possible correlation between DNA concentrations and voucher age (Fig. D 1.1.2).

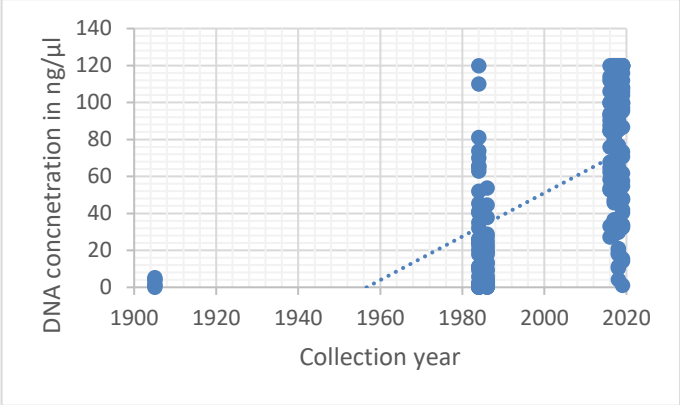

Fig. D1.1.2: DNA concentrations in 238 cichlid fishes dating from 1984 to 2019

However, after additional JEMU projects were initiated or finalised between 2022 and 2023, we shifted the focus towards two alternative and more informative datasets including (a) barn owls (*Tyto alba alba*) from both historical as well as more recent bird collections and (b) pollen recovered from pollinating Diptera and Hymenoptera from the insect collections of RMCA. These data, complemented by those collected in the framework of the activities of W2 and WP3, were considered for the validation of the results of this project (see D.3.2.1).

InsectMOoD largely relied on synergies and co-financing from the projects DISPEST, AGROVEG and DIPODIP (framework agreement 2019-2023, Royal Museum for Central Africa - Directorate-general Development Cooperation), FFI-PM (EU, H2020, grant 818184), REACT (EU, H2020, grant 101059523) and SYNTHESYS+ (EU, H2020, grant 823827). The datasets assembled in these projects included WGS data which contributed to the production of > 1,300 genomic vouchers from the target taxa (*Tephritidae* and *Syrphidae*) and which were archived in the collections of RMCA through InsectMOoD (see D.3.1.1.).

D 1.2.1: selection of suitable methods for testing (M6)

The datasets listed above were used for dedicated experiments on the performances of -omic approaches on Museum collections according to the methods detailed in D2.1.2.

WP2. Comparisons of WGS performances on insect Museum vouchers

Task 2.1. Experimental design and testing

Vouchers were selected from the collections of *Syrphidae* and *Tephritidae* of RMCA. The selection included both recent and historical samples, dried / pinned and preserved at room temperature or preserved in ethanol at -20 / -80 °C. Levels of DNA concentration and degradation were quantified. Replicated DNA extracts for each combination of insect family (*Syrphidae*, *Tephritidae*), DNA concentration, fragmentation and contamination were subjected to genomic library preparation and high throughput sequencing (HTS). The results obtained were analysed through a range of statistical approaches (including Generalised Linear Model) and the performance of HTS was compared across Museum samples with different features (see methods and results detailed in D.3.1.1 and D.3.1.2).

Task 2.2. A decision map for the routine genotyping of Diptera from Museum collections

The amount of base pairs /reads recovered, the proportion of high / low quality reads, as well as the cost and workload per sample (including time needed for both wet- and dry-lab procedures) were

estimated for the experimental samples of T 2.1. The most cost- and time-effective methodological approach for the routine genotyping of Diptera from Museum collections samples with different levels of DNA concentration and degradation was indicated in a decision map (see D.2.2.2). General recommendations were formulated in section 4 of this report).

D 2.1.1: project experimental setup (M6)

The general project experimental setup to characterise relationships between sample features in *Syrphidae* and *Tephritidae* vouchers, DNA quality ad quantity and WGS performance was preliminarily defined during the initial phases of this project (see year report 2021). Starting from M4, vouchers from the Diptera collections of RMCA were selected and processed and DNA quality metrics were recorded (see section below).

D 2.1.2: report on test for target taxa (M12)

D 2.1.2.1: Preliminary results 2021

Comparative performances of commercially available DNA extraction kits

Preliminary lab tests on *Syrphidae* and *Tephritidae* were completed during the first year of InsectMOoD. Figure D2.1.2.1.1 and D2.1.2.1.2 illustrate DNA concentrations and n. of HTS reads obtained in a subset of Museum specimens from the two target insect families.

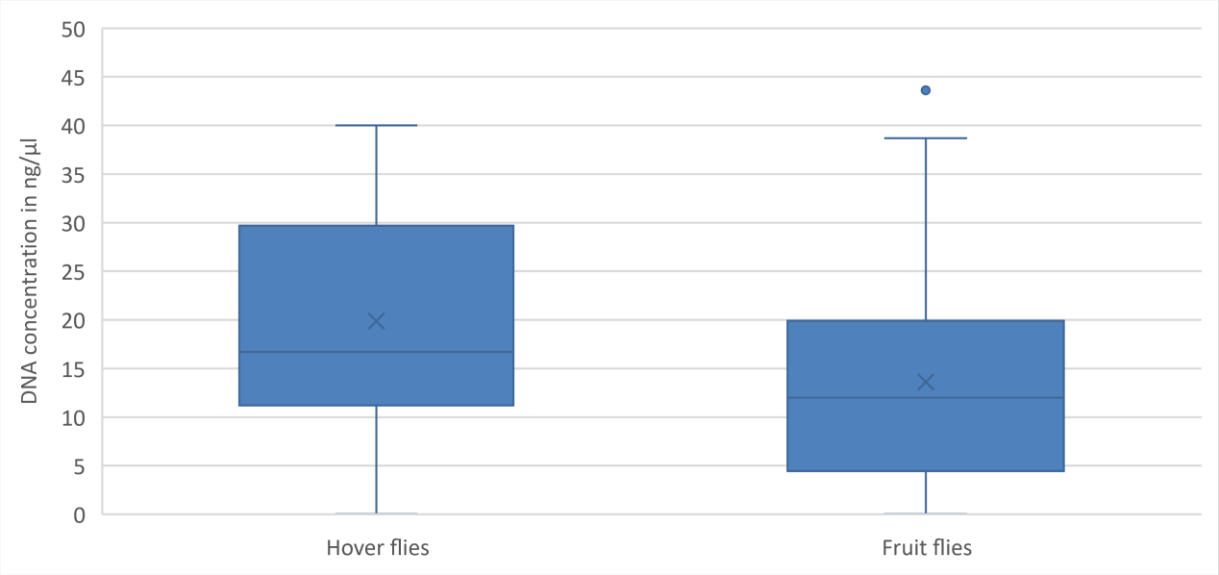

Figure D2.1.2.1.1 1: Boxplots of DNA concentration (ng/µl) in Syrphidae (n = 99) and Tephritidae (n = 149).

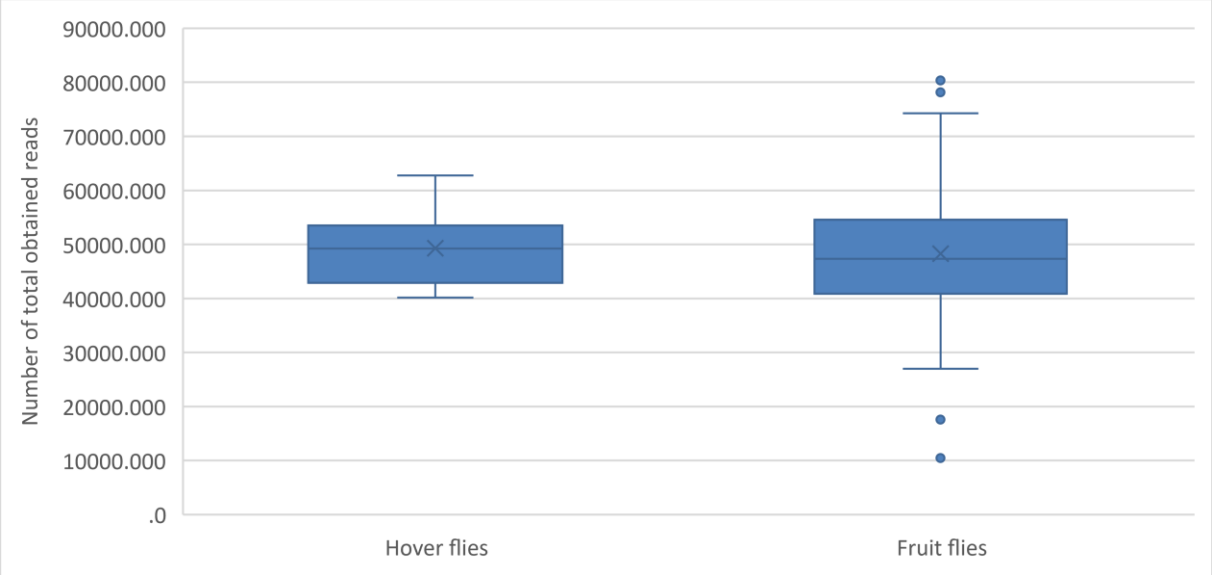

Figure D2.1.2.1.2: Boxplots of total n. of HTS reads obtained in Syrphidae (n = 99) and Tephritidae (n = 149).

The performance of commercial DNA extraction kits were compared in a pilot study targeting the RMCA collections of *Tephritidae* and *Syrphidae* (Diptera). We selected 3 to 6 specimens from seven collection series dating from 2008 to 2016. These included three *Tephritidae* (*Zeugodacus cucurbitae* (Coquillett), *Bactrocera dorsalis* (Hendel), *Dacus bivittatus* (Bigot)) and two *Syrphidae* (*Eumerus* sp. and *Ischiodon aegyptuis* (Wiedemann)) species, and specimens which were either stored in 100% ethanol at -20°C (Zc, Bd, Db, Eu) or pinned and preserved at room temperature (Ia). Digestions in lysis buffers were implemented on whole bodies for all specimens. For comparative purposes, we also processed forelegs only, rather than whole insects. The lysates obtained from each specimen were divided in four aliquots and the DNA purified using spin columns from the DNA extraction kits listed in Table D2.1.2.1.1 following the manufacturer's instructions. The experimental design was based on 30 whole specimens and 18 legs (2 negative controls were also included) which provided aliquoted DNA that was processed through 200 spin columns from four different extraction kits. The concentration of each DNA extract was measured using a Qubit 3 fluorometer (HS DNA Kit, Thermo Fisher Scientific) and the total amount of DNA was inferred from the final elution volume, which in all cases was 100 µl.

Table D2.1.2.1.1: Overview of the DNA extraction kits tested.

QIAGEN kit (50 samples)	**Column**	**Range DNA size**	**Expected DNA yield (according to manufacturer's instructions)**
DNeasy Blood and Tissue Kit	Dneasy spin column	100 bp-50 kb	6-30 µg
QIAamp Micro Kit	QIAamp MinElute column	<30 kb	<3 µg
QIAamp Mini Kit	QIAamp Mini spin column	<50 kb	4-30 µg
DNeasy Blood and Tissue Kit	MinElute column (MinElute PCR Purification Kit)	70 bp-4 kb	<5 µg

The figures show heterogeneous DNA yields across specimens with values ranging from 57.8 to 153.0 ng for whole bodies and for legs 1.3 to 22.0 ng (as expected, due to the lower amount of tissue in legs compared to whole bodies). An overview of the DNA yields is provided in Fig. D 2.1.2.1.3 and Fig. D 2.1.2.1.4.

The analysis of yields obtained from different DNA extraction methods shows that the kits tested provided comparatively similar performances, when used on representative specimens selected from our insect collections (Fig. D 2.1.2.1.3). Therefore, considering const-benefits we decided using the kit with the lowest price among those tested and routinely use the DNeasy columns in the routine processing of vouchers from the target insect collections.

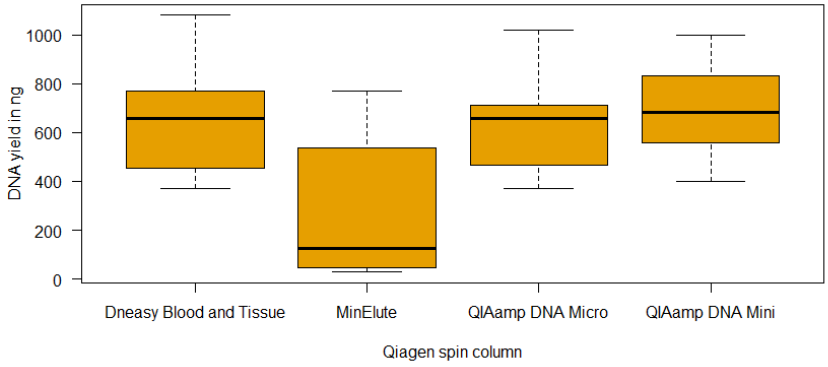

Figure D2.1.2.1.3: Boxplots of DNA yields from replicated elutions (whole-body digestions) per DNA extraction kit.

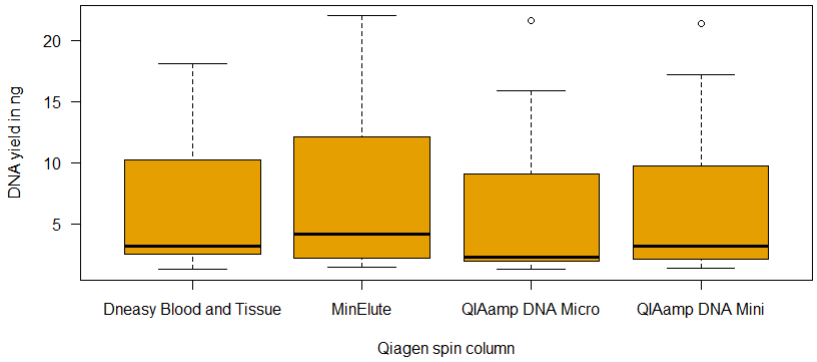

Figure D2.1.2.1.4 3: Boxplots of DNA yields from replicated elutions (leg digestions) per DNA extraction kit.

D 2.1.2.2: General results 2021-2022

During the second year of this project, the tests were implemented on a more extensive set of collection vouchers. The results of these analyses are currently being assembled in a draft publication to be submitted to a peer-reviewed journal. General conclusions and recommendations are reported in section 4 of this document.

Levels of DNA degradation in the RMCA insect collections were estimated in a selection of 1,405 vouchers from the two target Diptera families *Tephritidae* ("true" fruit flies), and *Syrphidae* (hoverflies or flower flies) from specimens collected between 1997 and 2022 in 54 countries in Africa (n = 925), America (n = 30), Asia (n = 83) and Europe (n = 367). The selection included 1,296 *Tephritidae* from 4 genera (*Bactrocera*, *Ceratitis*, *Dacus, Zeugodacus*) and 79 species and 109 *Syrphidae* from 2 genera (*Eristalinus* and *Melanostoma*) (Fig. D 2.1.2.2.1; Table D 2.1.2.2.1). We targeted these collections because they are very actively maintained and because of their relatively known sampling history (i.e. often including information about field collection methods, sample preservation protocols, habitat features, etc.).

A large proportion of the DNA samples were obtained from insect vouchers preserved either at -80°C (n = 671) or at -20°C in absolute ethanol (n = 653) while a minor proportion from pinned collections (n = 14). The DNA collections of RMCA, which are maintained through dedicated long-term stabilization protocols (see https://gentegra.com/gentegra-dna-2/) were also considered (n = 67).

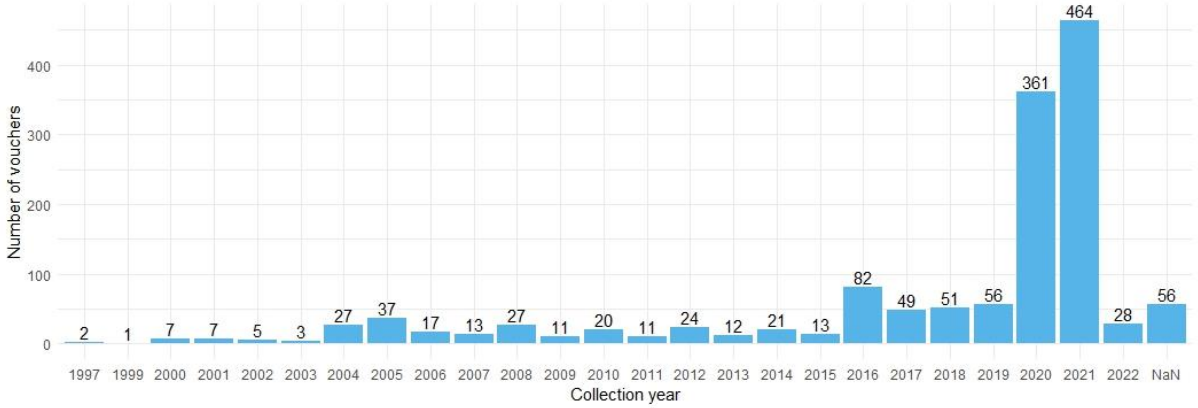

Figure D2.1.2.2.1 4: Year of sampling of the 1,405 collection vouchers considered in this study (number of vouchers / year is indicated).

Table D2.1.2.2.1 1: Regional sampling of the targeted vouchers (number of vouchers / country is indicated).

Region	Country	n
Africa	Anjouan	15
Africa	Benin	20
Africa	Burkina Faso	9
Africa	Burundi	13
Africa	Cameroon	5
Africa	Cape Verde	2
Africa	Central African Republic	1
Africa	Congo	9
Africa	Egypt	7
Africa	Ethiopia	12
Africa	Ghana	5
Africa	Grande Comore	14
Africa	Guinea	5
Africa	Ivory Coast	17
Africa	Kenya	34
Africa	La Réunion	1
Africa	Liberia	4
Africa	Madagascar	47
Africa	Malawi	14
Africa	Mali	1
Africa	Mauritius	23
Africa	Mayotte	20
Africa	Moheli	15
Africa	Mozambique	188
Africa	Nigeria	14
Africa	La Réunion	2
Africa	La Réunion	23

Region	Country	n
Africa	Senegal	6
Africa	Seychelles	1
Africa	South Africa	279
Africa	Sudan	11
Africa	Tanzania	45
Africa	Togo	24
Africa	Uganda	35
Africa	Zambia	3
Africa	Zimbabwe	1
America	Brazil	6
America	Costa Rica	7
America	El Salvador	5
America	Guatemala	6
America	Panama	6
Asia	India	18
Asia	Iran	38
Asia	Israel	7
Asia	Oman	6
Asia	Pakistan	7
Asia	Thailand	7
Europe	Austria	62
Europe	Croatia	107
Europe	Greece	14
Europe	Italy	156
Europe	Mayotte (France)	2
Europe	Spain	13
Europe	Switzerland	13

DNA was recovered from the targeted collection vouchers (n = 1,405) through non-destructive DNA extraction methods using the DNeasy Blood & Tissue Kit (Qiagen) with final eluted volumes of 60 µl (n = 140), 100 µl (n = 638) or 120 µl (n = 627). In most cases, the whole insect voucher was digested (n = 1,396). In a few cases (n = 9), abdomens (n = 8) or a leg (n = 1) were used.

The suitability of the extracted DNA for genetic / genomic research was estimated by quantifying:

- total DNA recovered and DNA concentration by using a Qubit 4 fluorometer (HS DNA Kit, Thermo Fisher Scientific).
- DNA fragment size distribution by using a fragment analyzer (Genomics Core – Leuven, DNF-930 dsDNA Reagent Kit, range 75 bp – 20000 bp).
- Sample DNA molarity (nm/l) and DNA molarity across fragment size ranges (< 350 bp; 351 – 1000 bp and 1001 – 180,000 bp) as estimated by the ProSize Data Analysis Software (v4.0.1.4; Agilent Technologies).
- levels of DNA contamination from different sources were estimated by measuring the absorbance ratios A260/280 and A260/230 with an Implen NanoPhotometer N60 Touch. The use of two different absorbance ratios aimed at detecting DNA contamination from different sources (Lucena-Aguilar *et al.*, 2016). To account for measure variability and increase accuracy, average absorbances were obtained from three replicated measures. Following Lucena-Aguilar *et al.* (2016), DNA was semi-quantitatively categorised as "pure" with 1.7 < A260/280 < 2.0 or with 1.8 < A260/230 < 2.2 or as "contaminated" with values deviating from these ranges.
- Most of these samples were subjected to Whole Genome Sequencing (WGS) at 10x coverage at Berry Genomics (n = 1,002) or Novogene (n = 304) on an Illumina NovaSeq platform (150 PE reads, 6Gb raw data output / sample). The suitability of the recovered HTS data for genomic research was estimated by the number of raw reads from the first sequencing run and by quantifying the proportion of high quality reads, as calculated by dividing the number of reads with Q score > 30 by the total number of raw reads. The raw reads were trimmed using the *fastp* tool (Chen et al., 2018) and mapped to a *Drosophila melanogaster* reference genome (GenBank accession GCA_029775095.1) using the *bwa-mem* command from the burrows-wheeler aligner tool (Li and Durbin, 2009). The proportion of raw reads aligned was calculated.

Relationship between voucher features and DNA and HTS data quality were verified through multiple linear regression. Voucher collection year and family were considered as independent variables, while (a) total DNA recovered, (b) molarity of short DNA fragments (< 350bp), (c) absorbance ratios, (d) proportion of quality reads, (e) number of raw reads and (f) percentage of reads aligned to GCA_029775095.1 were considered as dependent variables.

Hypotheses were tested using a generalised linear model (GLM)) using the lme4 package (Bates *et al.*, 2009) in R (version 4.2.0). Linear models (LM) were fitted for all variables, except for the absorbance ratios (c) where the binary data fitted a generalised linear model with a binomial family. Model fit was assessed and multiple model specifications were examined, including the inclusion of interaction terms and alternative functional forms. Analysis of Variance (ANOVA) was implemented for hypothesis testing. The assumptions of normality and homoscedasticity were met where required. including the inclusion of interaction terms and alternative functional forms. The assumptions of normality and homoscedasticity were met where required (Sthle and Wold, 1989).

The amount of DNA recovered from 1,405 vouchers varied between 0.00 and 8.83 µg. A significant interaction between insect family and collection year was observed for the total DNA / voucher. For *Tephritidae* we observed an increase in the amount of DNA recovered over years (positive association), while there is no significant year effect observed for *Syrphidae* (Fig. D2.1.2.2.2, Table D2.1.2.2.2, D2.1.2.2.3).

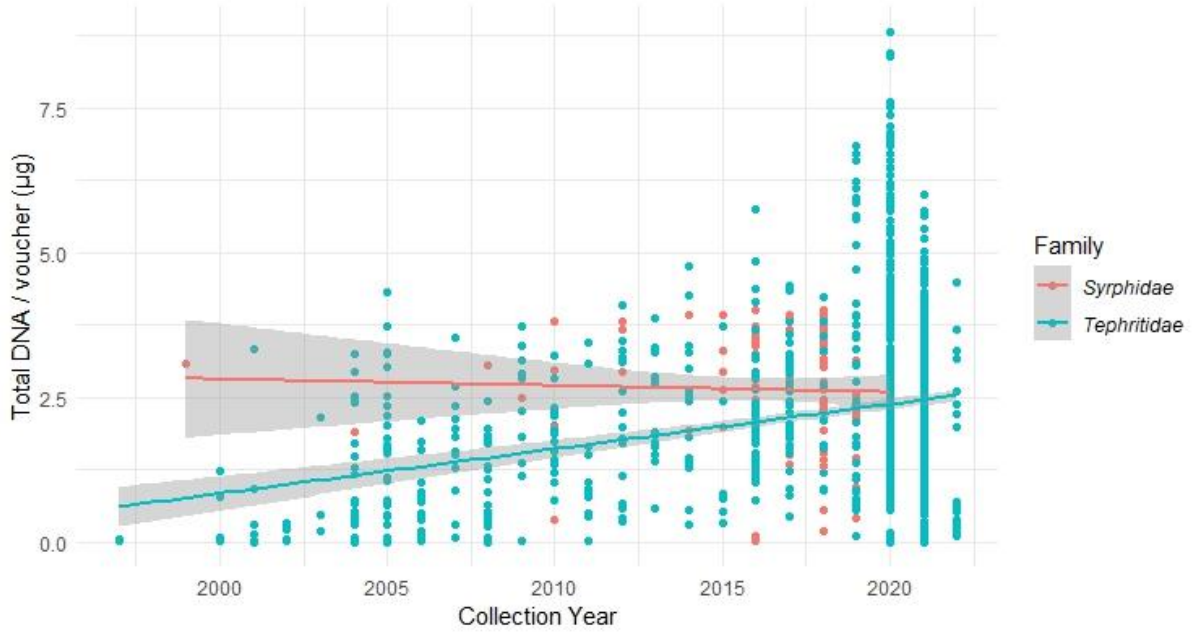

Figure D2.1.2.2.2: Total DNA / voucher (µg) in Tephritidae and Syrphidae.

*Table D2.1.2.2.2: Linear modelas (LM) and generalized linear models (GLM) testing for the effects of collection year and family on 7 variables. *: P<0.05, **: P<0.01, ***: P<0.001, n.s: non-significant. Detailed results of analyses of variance (ANOVAs) in Table D2.1.2.2.3.*

	AMOUNT OF DNA	PROPORTION MOLARITY (<350BP)	A280/260	A260/230	NUMBER OF RAW READS	PROPORTION QUALITY READS (Q>30)	PROPORTION ALIGNED READS
MODEL	LM	LM	GLM	GLM	LM	LM	LM
Collection year	***	***	**	***	n.s.	***	***
Family	***	***	***	***	*	***	n.s.
Collection year x Family	*	n.s.	*	n.s.	n.s.	***	n.s.

Table D2.1.2.2.3: Analysis of variance (ANOVA) of the (generalized) linear models testing for the effects of collection year and family on 7 variables.- P < 0.05 in bold. Df: degrees of freedom.

	Amount of DNA		Proportion of molarity (< 350bp)	
	F-value	*P*-value	*F*-value	*P*-value
Collection year	84.651	**< 2.2e-16**	30.548	**4.269e-08**
Family	13.099	**0.000306**	75.541	**< 2.2e-16**
Collection year x Family	3.858	**0.0497**		
	A260/280		A260/230	
	χ^2-value	*P*-value	χ^2-value	*P*-value
Collection year	107.224	**0.00105**	28.754	**8.22e-08**
Family	187.769	**1,47e-02**	70.895	**<2.2e-16**
Collection year x Family	38.499	**0.049750**		
	Number of raw reads		Proportion of Q > 30 reads	
	F-value	*P*-value	*F*-value	*P*-value
Collection year	0.9466	0.33079	242.714	**9,50e-04**
Family	46.124	**0.03193**	69.045	**0.0087044**
Collection year x Family			138.788	**0.0002038**

	Proportion of aligned reads (against GCA_029775095.1)	
	F-value	*P*-value
Collection year	185.233	**2,93e-02**
Family	21.365	0.1458

The proportion of short DNA fragments (nmol/l of fragments < 350 bp / nmol/l all fragments) varies between 0.542 and 1.000 (n = 902). The proportion of short DNA fragments differs significantly between *Tephritidae* and *Syrphidae* with lower values in *Tephritidae*, and a significant effect of collection year (Fig. D2.1.2.2.3; Table D2.1.2.2.2, D2.1.2.2.3). The patterns observed do not suggest obvious relationships between sample age and DNA fragmentation.

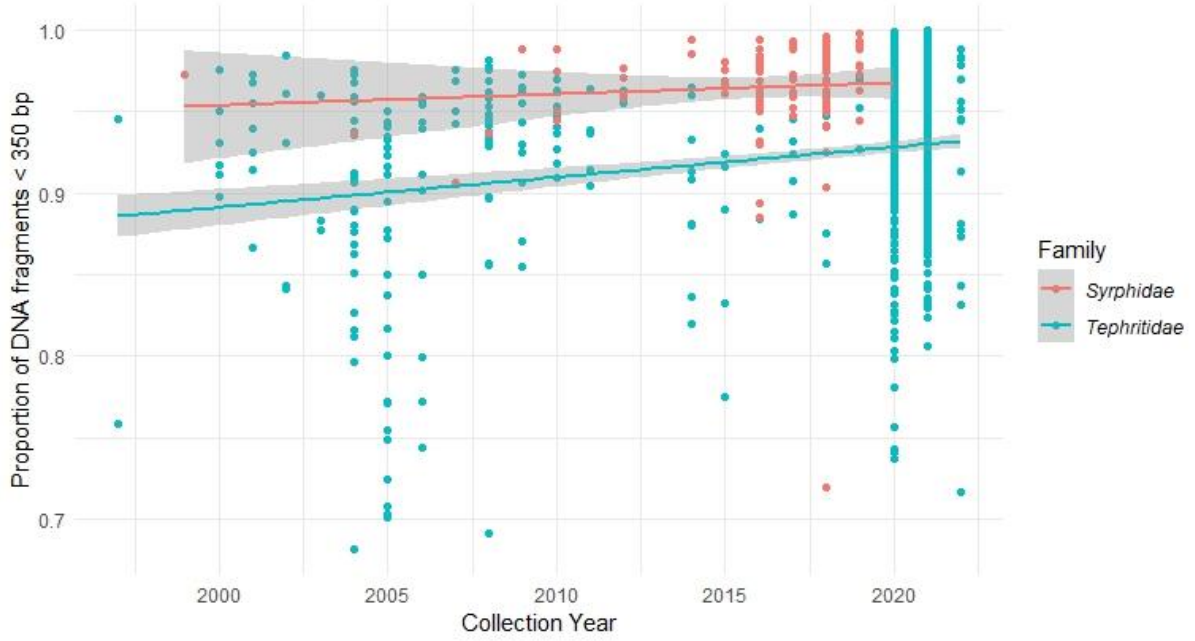

Figure D2.1.2.2.35: Proportion of short DNA fragments (<350bp) per collection year in Tehritidae and Syrphidae.

"Pure" DNA was observed in 16% of samples (out of 720) when considering A260/280 absorbance ratio and in 49% of samples when considering A260/230 ratios. A significant interaction between insect family and collection year was observed for the contamination patterns A260/280. For *Tephritidae* we observed an increase of the proportion of high quality DNA samples over years (positive association), while the opposite pattern (negative association) was observed in *Syrphidae.* Contamination patterns (A260/230)significantly differ between *Tephritidae* and *Syrphidae* and across years (Figure D2.1.2.2.4, D2.1.2.2.5, Table D2.1.2.2.2, D2.1.2.2.3).

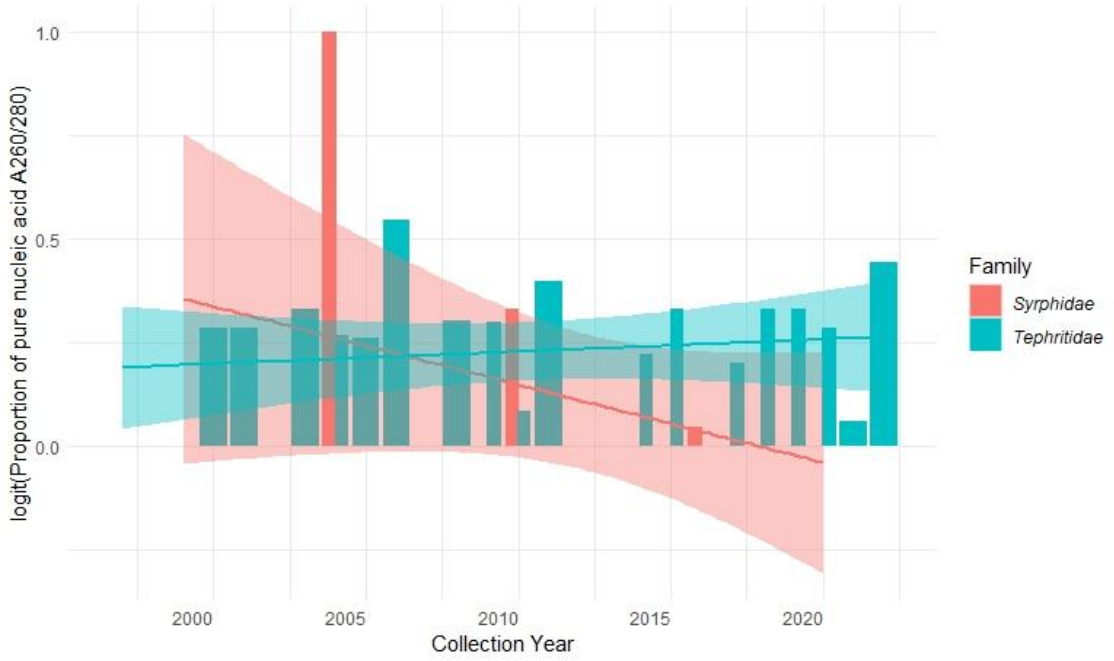

Figure D2.1.2.2.4 "Pure" A260/280 measures per collection year with separate regression lines per family.

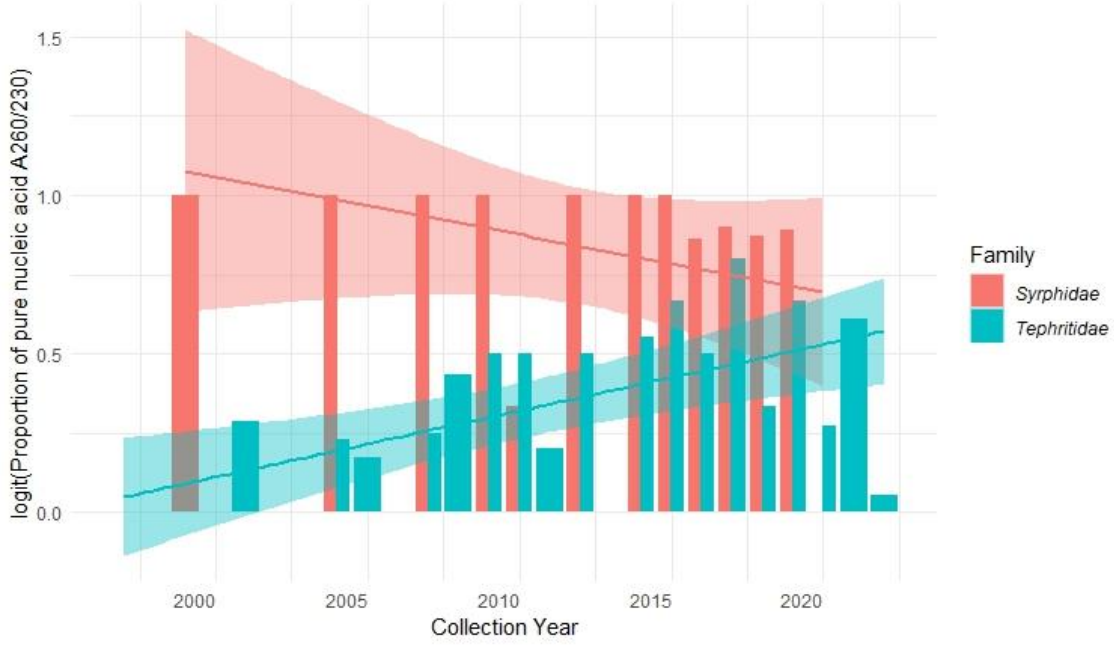

Figure D2.1.2.2.56: "Pure" A260/230 measures per collection year with separate regression lines per family.

The number of raw reads obtained from the first sequencing run from 1,304 samples varied between $0.16x10^7$ and $9.38\ x10^7$. The number of raw reads significantly differs between Tephritidae and Syrphidae, while we did not observe any significant effect of sample age (Fig. D2.1.2.2.6; Table D2.1.2.2.2, D2.1.2.2.3).

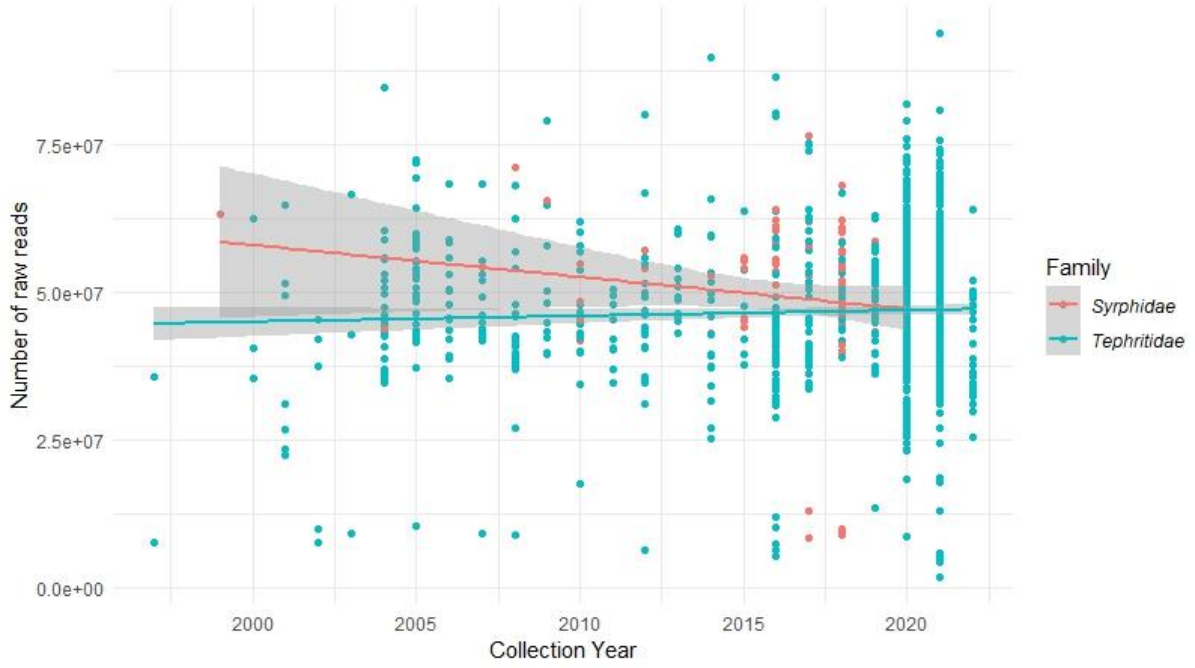

Figure D2.1.2.2.6: Raw reads per sample per collection year with separate regression line per family.

The proportion of high quality reads (i.e. with Phred Q score > 30) ranged from 0.8610 to 0.9554 in 1,305 samples. A significant interaction between insect family and collection year was observed. For *Syrphidae* we observed an increase of the proportion of high quality reads in recent samples, while the opposite pattern was observed in *Tephritidae* (Fig. D2.1.2.2.7; Table D2.1.2.2.2, D2.1.2.2.3).

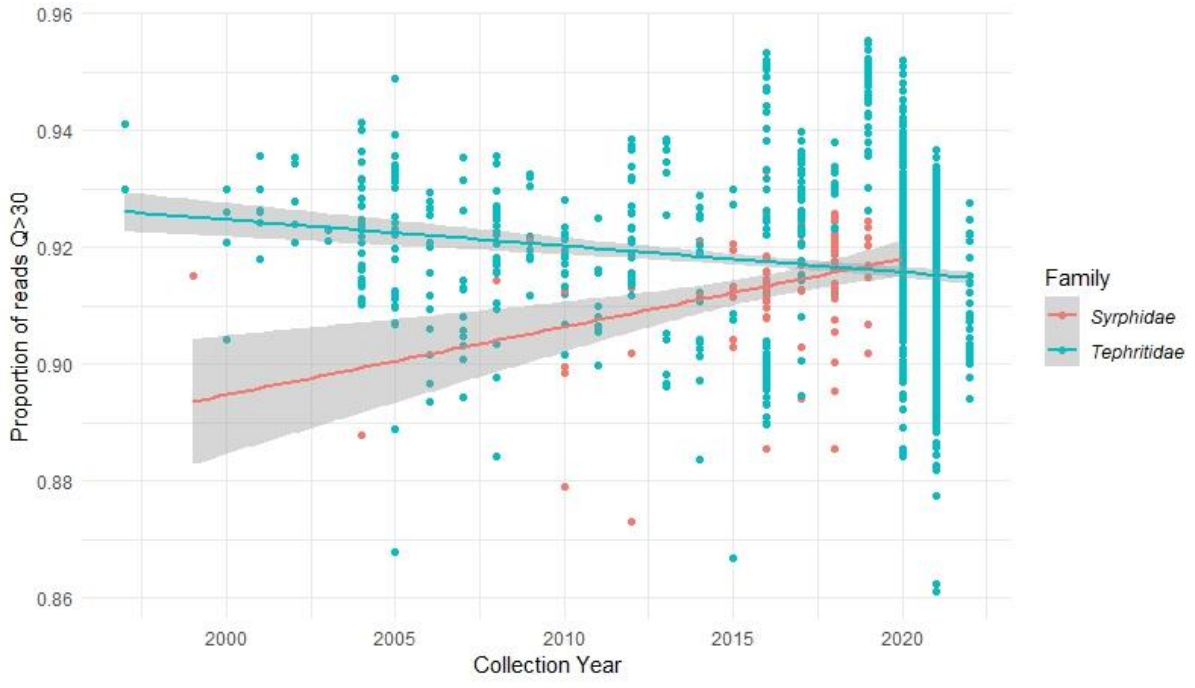

Figure D2.1.2.2.7: Pproportion of quality reads (Q>30) per collection year with separate regression line per family.

A subset of 162 sequenced DNA samples were aligned to a *Drosophila melanogaster* reference genome (GCA_029775095.1). The proportion of aligned reads ranged between 0.0025 and 0.2816. We did not observe significant differences between families, while a significant year effect was detected (Fig. D2.1.2.2.8; Table D2.1.2.2.2, D2.1.2.2.3).

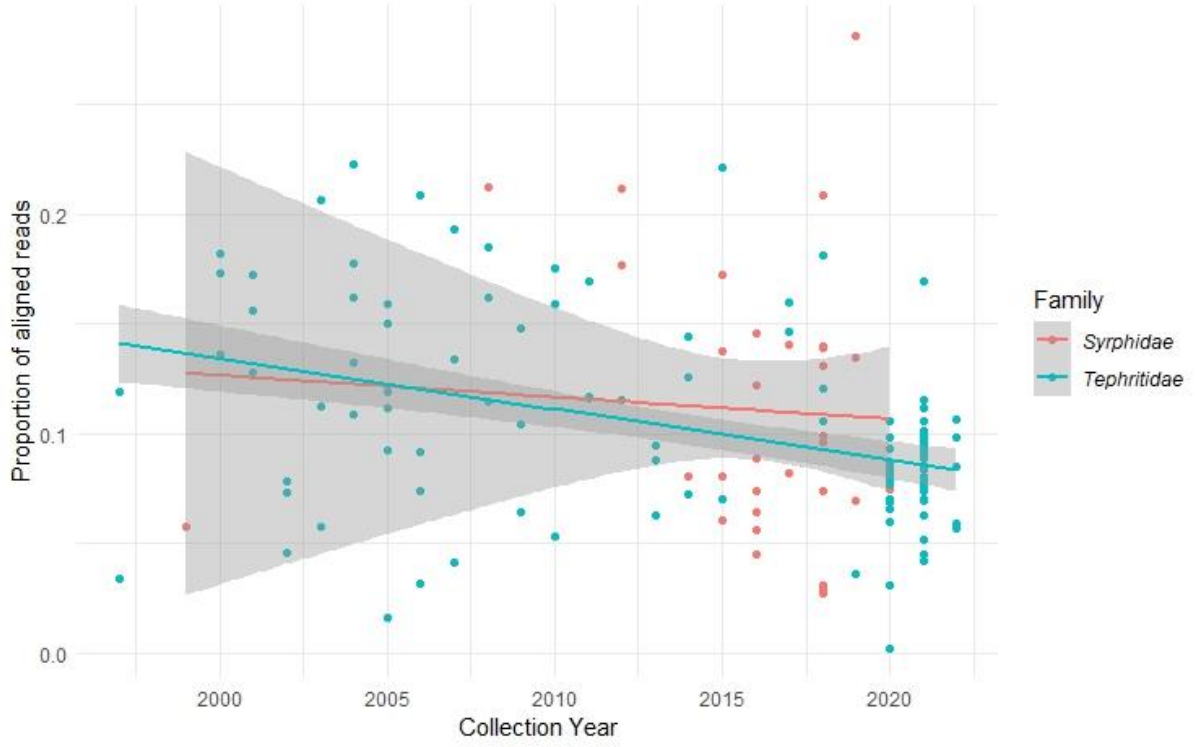

Figure D2.1.2.2.8: Proportion of reads aligned to GCA_029775095.1 per collection year with separate regression line per family.

D 2.2.1: cost-benefit analysis (M12)

A cost benefit analysis was implemented on DNA extraction procedures (4 commercial kits compared), the DNeasy kit was deemed to be the most cost-effective for the routine processing of the targeted samples from the collections of *Syrphidae* and *Tephritidae* (see D 2.1.2).

D 2.2.2: decision map for the WGS of suboptimal samples (M12)

A decision map for *Tephritidae* and *Syrphidae* from museum collections was provided during the first year of this project and is reported below (Fig. 11)

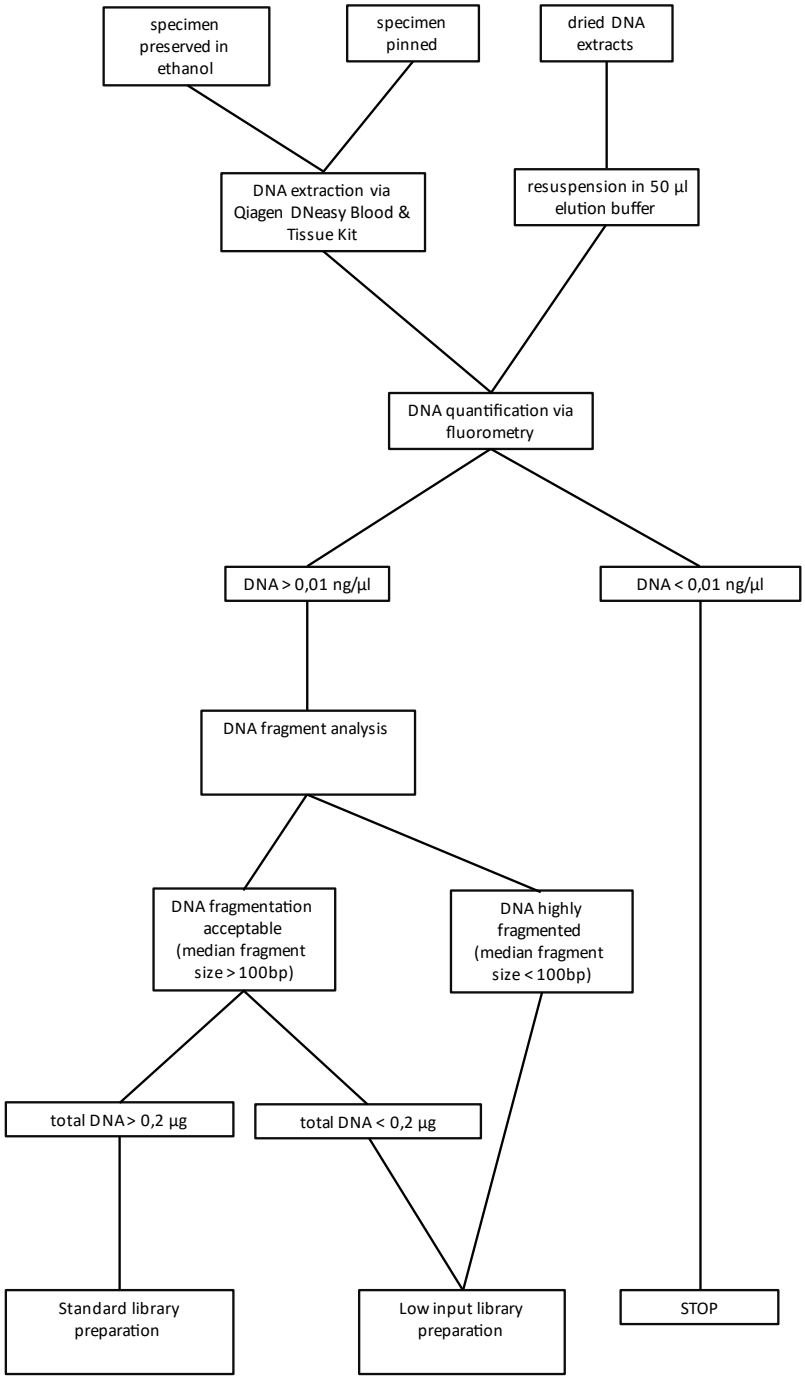

Figure 7: A decision map for the routine genotyping of Tephritidae and *Syrphidae from museum collections.*

WP3: production and archiving of genomic and digital vouchers

Task 3.1: Voucher genotyping

Based on the results of Task 2.2, we applied the most appropriate WGS protocol on insect vouchers from the RMCA collections giving priority to specimens of particular interest for the ongoing research lines on Diptera systematics or population genomics.

Task 3.2. Validation of results

To verify the generality of patterns observed for the two target Diptera families, we also explored relationships between DNA concentration / degradation and HTS performance in non-target groups of collection samples. In this respect, we used datasets collected in the framework of projects ongoing at RMCA/RBINS in 2021-2023 and verified to which extent relationships between DNA degradation and HTS performance observed in the target groups (see D1.1.2, D2.1.2.) could be extended to different groups of Museum samples (D3.2.1).

Task 3.3. Voucher archiving and open access

The long-term storage and archiving of the extracted DNA, based on routine protocols used by the JEMU, was finalised. A subset of vouchers was digitalised and archived using available resources and established procedures at RMCA. The genomic vouchers were incorporated in an open-access database after linkage to (a) voucher collection code, (b) voucher digital image, (c) DNA voucher code, (d) link to whole genome sequencing data. All metadata are publicly available via the website of RMCA, using standardised databasing procedures. Voucher DNA samples and genomic data will be available in the framework of national and international collaborative research following agreement with the partner Institutions. The links to whole genome sequencing data, provisionally represented by the path to the Network-attached storage (NAS) systems of RMCA, will be gradually replaced by links to public databases such as NCBI as soon as the data will be published in the framework of ongoing or future research projects.

D 3.1.1: report on voucher WGS and assemblage (M18, M24)

Synergies between InsectMOoD and the projects DISPEST, AGROVEG and DIPODIP (framework agreement 2019-2023, Royal Museum for Central Africa - Directorate-general Development Cooperation), FFI-PM (EU, H2020, grant 818184), REACT (EU, H2020, grant 101059523) and SYNTHESYS+ (EU, H2020, grant 823827) allowed collecting WGS data from > 1,300 *Tephritidae* and *Syrphidae* from different geographic regions and sampling years (D2.1.2.). Raw WGS reads were filtered and processed in the framework of the research activities planned by the projects listed above. All data were stored on the local servers of RMCA as detailed in D3.3.2 and D3.3.3).

D 3.2.1: report on tests for non-target taxa (M18, M24)

Tests on performances of reduced representation sequencing (RRS) of collection vouchers and metabarcoding of pollen recovered from insect collections were performed in synergy with JEMU internal research implemented between 2022 and 2023. The results of these tests are reported below.

D.3.2.1.1: Reduced-representation sequencing of museum collections.

Natural history collections have been brought forward as a valuable tool to contrast genomic patterns before and after an environmental change took place (Wandeler, Hoeck and Keller, 2007; Holmes *et al.*, 2016). As such, they grant us a window to the past and provide unparalleled strong statistical designs to address knowledge gaps (Card *et al.*, 2021). Identifying temporal trends often requires substantial sample sizes making whole-genome sequencing too costly for many projects. A variety of restriction digest derived methods have offered scientists a way out to obtain high-resolution population genomic data at reduced costs (Davey and Blaxter, 2010; Puritz *et al.*, 2014). All these highly related methods apply a restriction digest of the genome and subsequent sequencing of the outer ends of each digested fragment. Hence, only a fraction of the genome is sampled, e.g. reduced representation sequencing (RRS), but still thousands of genetic markers across all samples are obtained at reasonable costs (Davey and Blaxter, 2010). Although this technology has proven its value when working with high quality DNA (Nadeau *et al.*, 2014; Van Belleghem *et al.*, 2018), its

implementation in museum studies has been vastly hampered by the unpredictable outcome due to DNA degradation (Graham *et al.*, 2015; Souza *et al.*, 2017; Lang *et al.*, 2020). DNA fragmentation at restriction sites causes failure or bias in RRS by inefficient or lack of restriction digest, while random shearing lowers the number of fragments being flanked by both restriction sites and therefore prevents the necessary adapter ligation (Puritz *et al.*, 2014; Graham *et al.*, 2015). A study on artificially induced DNA degradation illustrated a significant decrease in the number of RADtags per individual, number of variable sites, and percentage of identical RADtags retained (Graham *et al.*, 2015). These difficulties have refrained scientists from using RRS as a tool to obtain museum population-level data. Yet, when large collections are available, a careful screening assessment prior to the onset of library preparation could aid in narrowing the focus exclusively on those samples exhibiting the greatest likelihood of success, rendering RSS still as a plausible cost-efficient tool to extract population-level data from historical collections.

Therefore, we here attempt to assess i) to what extent DNA degradation affects the success rate of RRS in a long term time series of avian museum studies, and ii) whether we can predict *a priori* the success rate of RRS on museum samples using easily to obtain DNA quality metrics.

We sampled 96 barn owls (*Tyto alba alba*) comprising both historical as well as contemporary specimens. Historical samples were obtained from collections stored at the Royal Belgian Institute of Natural Sciences and covered two distinct periods in time, the 30's (1929-1943, n=15) and the 70's (1966-1979, n=22). Contemporary specimens (n=59) comprised road kills which were brought to bird sanctuaries and stored in freezers immediately upon arrival. We collected toe pads of all historical specimens to minimize voucher damage, and liver or breast muscle tissue of the contemporary specimens.

DNA of all specimens was extracted using the NucleoSpin tissue kit (Macherey-Nagel GmbH). Concentrations were quantified by the Qubit fluorometer (Invitrogen) and a fragment analysis of historical samples was conducted on a 2100 Bioanalyzer (Agilent). While numerous variations on reduced representation genome sequencing exist (Puritz *et al.*, 2014), we here focussed on a technology called double-digest restriction site-associated DNA sequencing (ddRAD) because of the simplified wet-lab workflow, low cost and highly homogenous coverage of sites across samples (Peterson *et al.*, 2012). DdRAD libraries were constructed following the protocol of Peterson *et al*. (2012). Briefly, we digested DNA samples using two restriction enzymes, i.e. SbfI and MseI. Starting volumes of DNA were adjusted according to sample specific DNA concentrations (18µl, 12µl or 6µl of DNA when concentrations were respectively lower than 20 ng/µl, between 20-32 ng/µl or higher than 32 ng/µl). Barcoded SbfI and universal MseI-compatible adapters were subsequently ligated to the digested genome, followed by a size selection of fragments of 270 bp ("narrow peak" setting) on a BluePippin (Sage Science). Lastly, fragments were PCR amplified using a barcoded forward primer to obtain dual-indexed ddRAD libraries, which were subsequently pair-end sequenced on an Illumina Novaseq6000 platform. Raw data were demultiplexed using the process_radtags module in Stacks v2.50(Catchen *et al.*, 2011). Trimmomatic v0.39 (Bolger, Lohse and Usadel, 2014) was used to remove adapters and a sliding window approach was applied to trim reads when quality fell below 20. Paired reads were mapped to a reference genome (GCA_000687205.1_ASM68720v1) using BWA mem (Li and Durbin, 2009) using default settings and only properly paired reads with a quality > 30 were retained using SAMtools v1.11 (Li *et al.*, 2009). SNPs were subsequently called using GATK's HaplotypeCaller tool (McKenna *et al.*, 2010).

In order to avoid any bias in downstream analyses arising from contaminated historical specimens, we first assembled a stringently filtered vcf based exclusively on recent samples. Specimens showing more than 20% missing data were discarded and only biallelic SNPs (--max-alleles 2) with a minimal SNP quality (--minQ) of 40 and an individual genotype (--minGQ) quality of 30, present in at least 50% of the individuals (--max-missing) and a minimum allele frequency (--maf) of 0.01 were retained with VCFtools (Danecek *et al.*, 2011). This resulted in a data set of 31012 SNPs. These reference SNPs were

then subsequently extracted from all individuals, e.g. both historical as well as contemporary specimens, to limit the erroneous inclusion of exogenous DNA sequences from historical samples.

We ran a one-way ANOVA to test for difference in mean number of missing SNPs between the three time periods, and allowed for period-specific variances to account for heteroscedasticity using the R package 'nlme' (version 3.1-160). To predict the success rate of ddRAD in museum samples we applied generalized additive models (GAM) to relate percentage of missing SNPs per individual to either DNA concentration or fragmentation using the R package 'mgcv' (Wood, 2011). All statistical analyses were performed using the R 4.1.2. software (R Core Team 2021). DNA fragmentation was assessed from Bioanalyzer profiles by calculating the percentage of the total area under the curve in four distinct bins, e.g. bins that contain fragments ranging from respectively 35-200bp, 200-400bp, 400-700bp or 700-10380bp.

Mean missing data per individual differed significantly between periods (χ_2=62.56, p<0.001) (Figure 1). The mean percentage of missing SNPs was 2.6% for recent specimens, 43.4% for specimens sampled at the 70s and 85.4% for specimens originating from 30s. The variance in missing data varied significantly between periods (Breusch-Pagan test, X^2=52.1, p<0.001). Recent samples showed consistently few missing SNPs, while the success rate in samples of the 30s varied slightly more. In contrast, samples of the 70s showed large variation in missing data, ranging from highly successful samples to those that failed almost completely, complicating the utility of age of the sample as a suitable predictor for success of RRS of museum specimens.

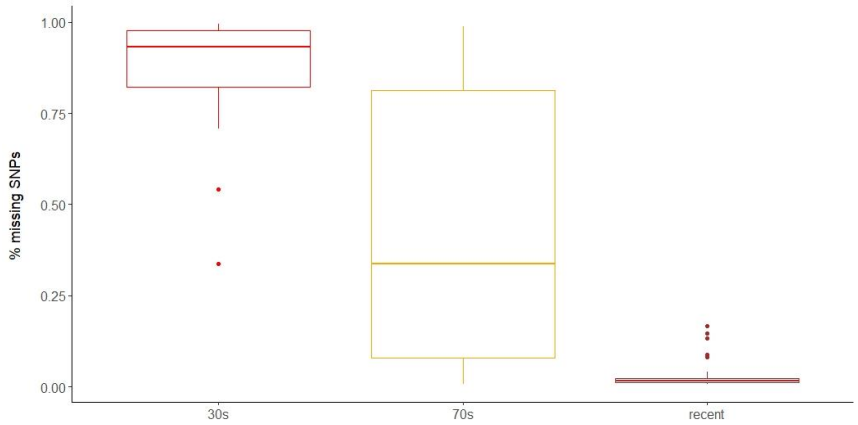

Figure D.3.2.1.1: Percentage of missing data per individual for each sampling period

Mean DNA concentration in historical and recent samples were respectively 20.2 ng/µl ± 12.4 (SD) and 30.6 ng/µl ± 13.9 (SD). A simple linear regression indicated the number of missing SNPs was not related to DNA concentration in recent samples ($F_{1,57}$=0.016, p=0.90). In contrast, a GAM indicated DNA concentration was inversely related to the amount of missing data in historical samples ($F_{1,3.2}$=15.97, p<0.001) and explained 66.8% of the deviance (Figure D.3.2.1.1).

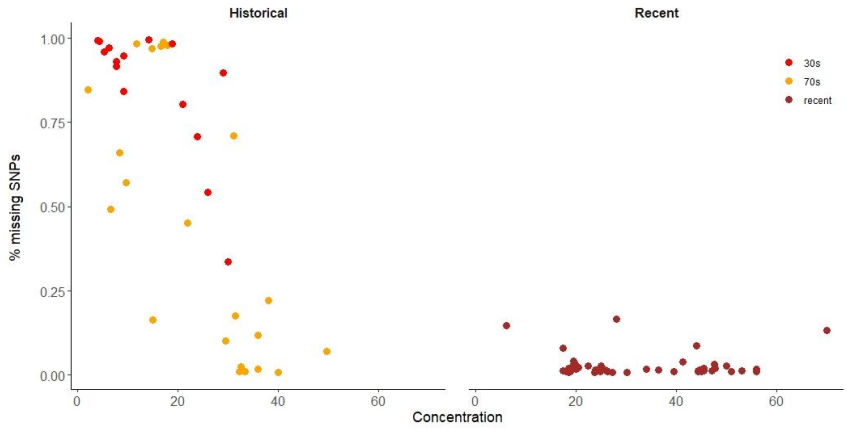

Figure D.3.2.1.2: Graphs depict the association between DNA concentration and percentage of missing SNPs in historical samples.

GAM's relating the amount of missing data to the percentage of fragments between 35-200bp, 200-400bp, 400-700bp and 700-10380bp explained respectively 74.8%, 20.7%, 39.7% and 78.4% of the model deviance. The amount of fragments in the lowest bin range was strongly positively associated to missing data ($F_{1,2.3}$= 32.99, p<0.001), while those at the highest bin range showed a clear negative association ($F_{1,2.4}$= 37.63, p<0.001) (Figure D.3.2.1.2). Based on the latter model the predicted amount of missing data when 1%, 5%, 10%, 20%, 30% or 50% percentage of fragments ranged between 700bp and 10380bp was respectively 88%, 77%, 65%, 42%, 23% *and 4%.*

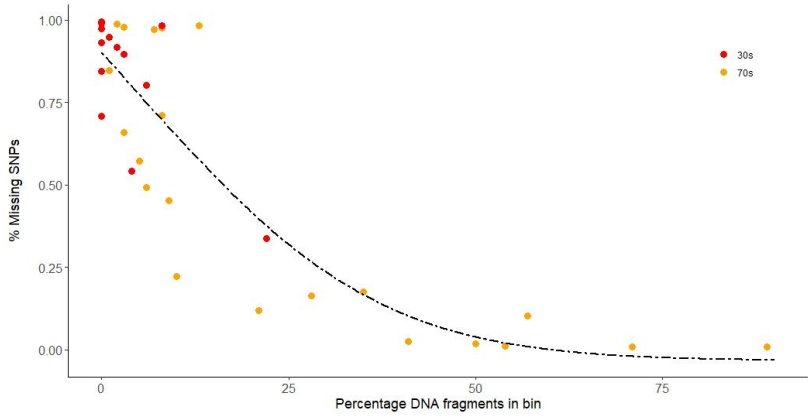

Figure D.3.2.1.3: Inverse association between fragmentation (i.e. percentage of fragments between 700 and 10380bp) and percentage of missing SNPs in historical samples. Dashed line represents the predicted values according to the fitted GAM.

To date, little guidance is provided on whether it is feasible, and if so how, to optimize museum sample selection at the onset of a RRS study given the unpredictable outcome when based on poor quality DNA (Souza *et al.*, 2017; Lang *et al.*, 2020). In a previous study using ddRAD target enriched sequencing it has been suggested to set the inclusion threshold of samples at a DNA concentration (as determined from the A_{260} values) of 30 ng/µl (Souza *et al.*, 2017). This value largely corroborates with our results (Fig. D.3.2.1.3), yet we advocate to rely on fragmentation assessments for several reasons. Firstly, DNA fragmentation was a better predictor for missing SNPs compared to DNA concentration. Secondly, DNA concentration was not always perfectly inversely associated to DNA fragmentation as some samples of low DNA concentration also showed low DNA fragmentation, or conversely, samples of high DNA concentration that were highly fragmented. DNA concentration as a decision metric would have resulted respectively in rejecting appropriate samples and including inappropriate ones. Thirdly, DNA concentration of problematic samples can be increased by eluting in smaller volumes or lysing more tissue during DNA extractions. Yet, fragmentation profiles will still remain unaffected and hence DNA molecules will still miss the appropriate restriction sites. Lastly, unlike fragmentation profiles, sample DNA concentrations are species and tissue dependent, making it unrealistic to set a universal threshold.

DdRAD appears unsuitable to obtain sequence data from highly fragmented samples, and more advanced target-capture based technologies (HyRAD, HyRAD-X) should be considered as an alternative when no other samples are available (Suchan *et al.*, 2016; Schmid *et al.*, 2017). These technologies unfortunately demand a high level of molecular expertise and are far less cost-efficient compared to ddRAD. However, obtaining population-level genomic data of museum specimens using ddRAD may still remain feasible when sufficiently large museum collections are available. Prioritizing samples based on fragmentation profiles will limit the focus on the most promising ones, and obtain high-quality data of sufficient samples in a cost-efficient manner.

(Also see general recommendations in section 4.2)

D.3.2.1.2.: Metabarcoding identification of pollen recovered from insect collections

Protocols for the metabarcoding identification (ID) of pollen loads from pollinating *Syrphidae* (Diptera) and Hymenoptera, Apoidea from the collections of the RMCA were optimised in synergy with the internal JEMU project POLBEN (2022). This project aimed at:

- surveying the literature to compare the suitability of different field and lab protocols available for pollen preservation and DNA extraction.
- testing a range of primers and PCR conditions for the Sanger sequencing of plant DNA barcodes, with a main focus to the PCR amplification of plants DNA barcodes from the family Cucurbitaceae, as a main target of ongoing research projects at RMCA.
- optimizing and comparing commercial and “in-house” protocols for preparing metabarcoding libraries
- verifying cost-benefits of outsourcing metabarcoding library prep.

Literature survey

A selection of references from the preliminary literature survey is reported in the end of this section.

Tests on optimal sample preservation and DNA extraction

In this test, we focused on pollen grains isolated from the bodies of “fresh” flower flies (Diptera, *Syrphidae*) collected by hand net in the Tervuren park, stored in 1.5ml Eppendorf tubes containing either 100% ethanol or 1ml of CTAB and subjected to:

a) Immediate pollen isolation and DNA extraction,
b) preservation for one month at room temperature (RT, ~25°C) followed by pollen isolation and DNA extraction,
c) preservation for one month in a freezer at -20°C followed by pollen isolation and DNA extraction (Tab. D.3.2.1.4.),

Pollen DNA was extracted using either a modified CTAB extraction protocol (Annex 2), elution volume 30ul) or the Qiagen DNeasy Plant Minikit (elution volume 100ul).

Table D.3.2.1.1 2: Test on sample preservation and pollen DNA extraction. Experimental setup and n. of specimens processed.

		Preservation group	CTAB DNA extraction	QIAGEN DNA extraction
Control (immediate processing after collection)		a	5	5
EtOH preservation (30 days)	Room T°	b	5	5
	Freezer (-20°C)	c	5	5
CTAB preservation (30 days)	Room T°	d	5	5
	Freezer (-20°C)	e	5	5

Pollen was isolated from the insect bodies by shaking the tube twice for 5 minutes on a bead beater at 6Hz (as a compromise between vigorously shacking the pollen and avoid damaging the vouchers).

The fly was then removed from a solution of 100% EtOH and the pollen centrifuged at 13,000 rpm for 5 min. The pellet was dried in a Eppendorf® Concentrator (1400 rpm for 60 mins) and disrupted in a bead beater (VWR® Star Beater) for 2 minutes at 22,5 Hz with three 3mm stainless steel beads per tube.

DNA concentration was quantified using a fluorometer (Qubit 3, HS DNA Kit, Thermofisher Scientific, Carlsbad, CA, USA). ANOVA and the Student-Newman-Keuls (SNK) as implemented by the R package GAD (Sandrini-Neto and Camargo, 2015), were used for *a priori* and *a posteriori* hypothesis testing with Extraction Method (CTAB vs EtOH) and Preservation Group (CTAB -20°C, CTAB RT, EtOH -20°C, EtOH RT, control) as fixed, orthogonal factors. As elution volumes were different in the two extraction methods (100ul for Qiagen, 30ul for CTAB), analyses were repeated for both total DNA yields and DNA concentrations.

Total DNA concentrations measured ranged from 0.020 ng/µl to 8.52 ng/µl and DNA yields from to 0.91ng to 284 ng (a minor part of measures resulted below the instrument detection limits).

ANOVAs on DNA yields and DNA concentration showed a significant interaction between insect preservation protocol and pollen DNA extraction method. *A posteriori* comparisons revealed significantly higher DNA yields concentrations for samples preserved in CTAB and subjected to CTAB DNA extraction (Fig. D.3.2.1.4., D.3.2.1.5., Annex 2).

This pattern could be artefactual as biased by cross contamination between pollen and insect DNA (with CTAB preservation promoting somehow insect DNA extraction?). In order to further explore this hypothesis, we tentatively amplified the DNA extracted by using PCR primers for both animal and plant DNA barcoding (Folmer *et al.*, 1994; Newmaster, Fazekas and Ragupathy, 2006). The comparative analysis of amplification success provided a first indication that CTAB preservation followed by CTAB DNA extraction (irrespectively of preservation temperature) seems to favour cross-contamination between plant and insect DNA. In fact, primers for animal DNA barcoding generally provide positive amplification on these samples (Tab. D.3.2.1.2).

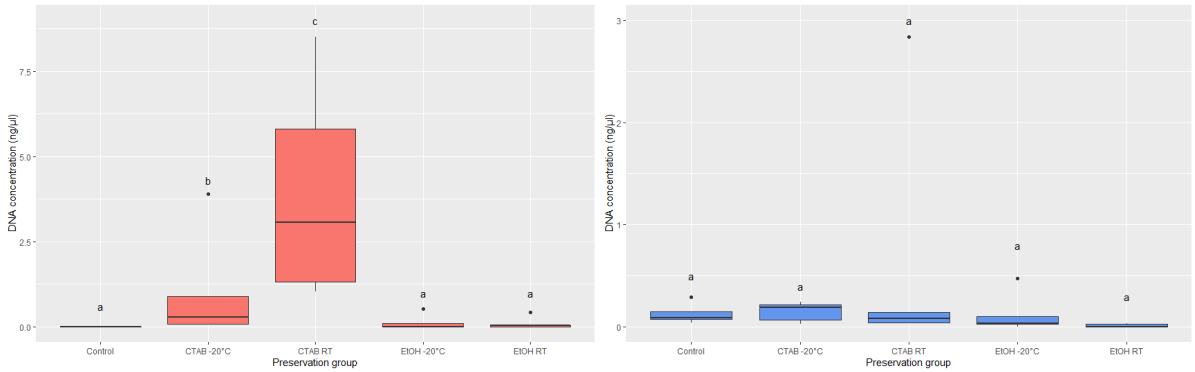

Figure D.3.2.1.4.8: DNA concentrations of pollen samples as recovered by using either the Qiagen DNeasy Plant Mini Kit (left) or CTAB DNA extraction (right). Pollen was either extracted immediately after collection (control) or preserved for 30 days in ethanol 100% at room temperature (EtOH RT); at -20°C (EtOH -20°C), in CTAB at room temperature (CTAB RT) or at -20°C (CTAB -20°C).

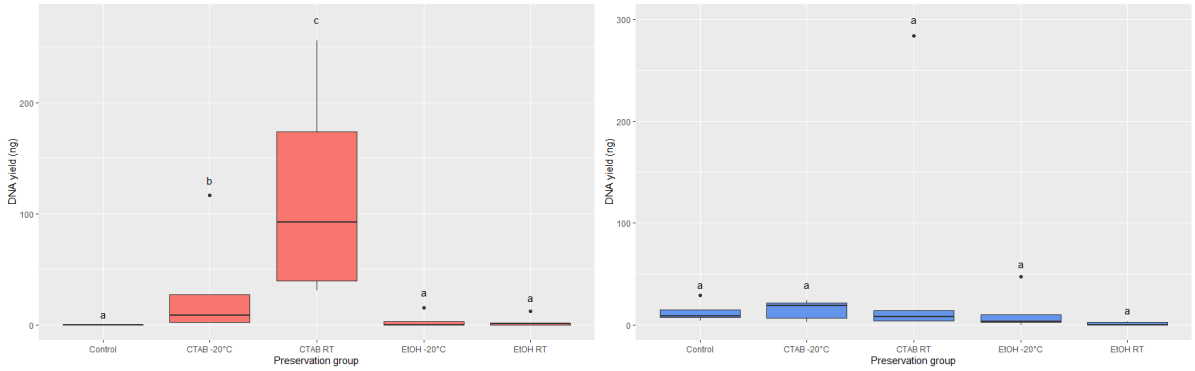

Figure D.3.2.1.59: Total DNA yield of pollen samples as recovered by using either the Qiagen DNeasy Plant Mini Kit (left) or CTAB DNA extraction (right). Pollen was either extracted immediately after collection (control) or preserved for 30 days in ethanol 100% at room temperature (EtOH RT); at -20°C (EtOH -20°C), in CTAB at room temperature (CTAB RT) or at -20°C (CTAB -20°C).

Table D.3.2.1.2: Exploratory, semi-quantitative test for contamination of pollen samples. Comparative PCR amplification (+ = amplification, - = no amplification) of pollen samples from different preservation groups (see D.3.2.1.1) via universal primers for plant (rbcl) and animal DNA (COI) barcoding. Positive COI amplification suggests contamination from insect DNA.

DNA extraction	preservation group			rbcl (plant)	COI (animal)
Qiagen	b	EtOH	RT	+	+
Qiagen	b	EtOH	RT	+	-
Qiagen	b	EtOH	RT	+	-
Qiagen	b	EtOH	RT	+	-
CTAB	d	CTAB	RT	+	+
CTAB	d	CTAB	RT	+	+
CTAB	d	CTAB	RT	+	+
CTAB	d	CTAB	RT	+	-
CTAB	e	CTAB	-20°C	+	+
CTAB	e	CTAB	-20°C	+	+
CTAB	e	CTAB	-20°C	+	+
CTAB	e	CTAB	-20°C	+	-

D.3.2.1.2..3: *Optimization of primers and PCR conditions for Sanger sequencing*

Wet-lab pipelines for the amplification of 4 markers generally used in plant DNA barcoding ID were developed: internal transcribed spacer 1 and 2 (ITS1 and ITS2), ribulose 1,5-biphosphate carboxylase (rbcL), and maturase K (matK).

Fourteen primer pairs were tested. Four of them were designed *ex novo* using Primer3 (Untergasser *et al.*, 2012) on alignments including 1500-2100 publicly available plant sequences (focus on *Cucurbitaceae*, minimum sequence length 500bp). Gradient PCRs (55°C < T < 65°C) were used to test optimal annealing temperature (Ta, Annex 2) on DNA extracts from three cucurbits (*Cucumis sativus* L.), pumpkins (*Cucurbita maxima* Duchesne), watermelon (*Citrullus lanatus* (Thunb.) Matsumura & Nakai)).

PCR were performed in a final volume of 25µl using the Platinum™ Taq DNA Polymerase (Invitrogen™). The PCR reaction mixture contained 2,50µl of PCR Buffer 10x, 0,75µl MgCl Platinum™ 50mM, 2,50µl of dNTP 2mM, 0,5µl of the forward primer (20µM) and 0,5µl of the reverse primer (20µM), 0,15µl Taq Platinum™ (5U/µl). PCR cycles included an initial heat activation for 5 min at 94°C; followed by 40 cycles of 30 s at 94°C, 30 s at Ta (see Tab. D.3.2.1.3:), and 1 min at 72°C; followed by a final extension of 10 min at 72°C.

Table D.3.2.1.3: Primer list for ITS1, ITS2, rbcL and matk DNA barcodes. The expected amplicon size was inferred using a selection of plant DNA sequences downloaded from the NCBI reference database.

gene fragment	Primer pair ID	Primer Forward	Primer Reverse	Expected amplicon size (bp)	Reference
ITS1	ITS1-390	AGTCGTAACAAGGTTTCCGT	GGGATTCTGCAATTCACACC	390	J. Ody – RMCA, unpublished
	ITS1-380	AGTCGTAACAAGGTTTCCGT	AACTTGCGTTCAAAGACTCG	380	J. Ody – RMCA, unpublished
ITS2	ITS2-23	ATGCGATACTTGGTGTGAAT	GACGCTTCTCCAGACTACAAT	460	(Chen *et al.*, 2010)
	ITS2-34	GCATCGATGAAGAACGCAGC	TCCTCCGCTTATTGATATGC	350	(White *et al.*, 1990)
	ITS2-54	CCTTATCATTTAGAGGAAGGAG	TCCTCCGCTTATTGATATGC	750	(Chen *et al.*, 2010)
	ITS2-Uni	TGTGAATTGCARRATYCMG	CCCGHYTGAYYTGRGGTCDC	310	(Moorhouse-Gann *et al.*, 2018)
rbcL	rbcL-506	ATGTCACCACAAACAGAGACT	AGGGGACGACCATACTTGTTCA	506	Modified from De Vere *et al.*, 2012
	rbcL-375	ATGTCACCACAAACAGAGACT	ACCCACAATGGAAGTAAACATGT	375	J. Ody – RMCA, unpublished
	rbcL-320	ATGTCACCACAAACAGAGACT	GCAAATCCTCCAGACGTAGA	320	J. Ody – RMCA, unpublished
	rbcL-23506	CTTACCAGYCTTGATCGTTACAAAGG	AGGGGACGACCATACTTGTTCA	275	(De Vere *et al.*, 2012; García-Robledo *et al.*, 2013)
	rbcL-T	ATGTCACCACAAACAGAGACT	GAAACGGTCTCTCCAACGCAT	660	Modified from Gous *et al.*, 2019
	rbcL-2623	CCTTTGTAACGATCAAGRCTTGGTAAG	CTTACCAGYCTTGATCGTTACAAAGG	380	(García-Robledo *et al.*, 2013)
	rbcL-A	ATGTCACCACAAACAGAGACTAAAGC	CTTCTGCTACAAATAAGAATCGATCTC	600	(Kress and Erickson, 2007)
matK	KIM	CGTACAGTACTTTTGTGTTTACGAG	ACCCAGTCCATCTGGAAATCTTGGTTC	890	(Laha *et al.*, 2017)

The performance of primers at suboptimal DNA concentrations were explored by diluting the three plant DNA extracts (one for each of the target cucurbits) to 0,1 - 0,05 - 0,01 - 0,001 - 0,0001 ng/µl and by verifying their amplification success. It's important to notice that concentrations below 0,05 ng/µl are below the detection limits of most fluorometers. Part of the 210 PCR products obtained (approx. 6%) were sequenced (Macrogen) to verify the amplification of the target gene fragment and exclude potential amplification / sequencing issues.

Five primer pairs (ITS1-390, ITS1-380, rbcL-506, rbcL-320, rbcL-23506) worked also at the lowest concentrations, rbcl-A did not provide any PCR product and was discarded from further consideration, the other 8 primer pairs generally worked at higher DNA concentrations (Fig. D.3.2.1.6.).

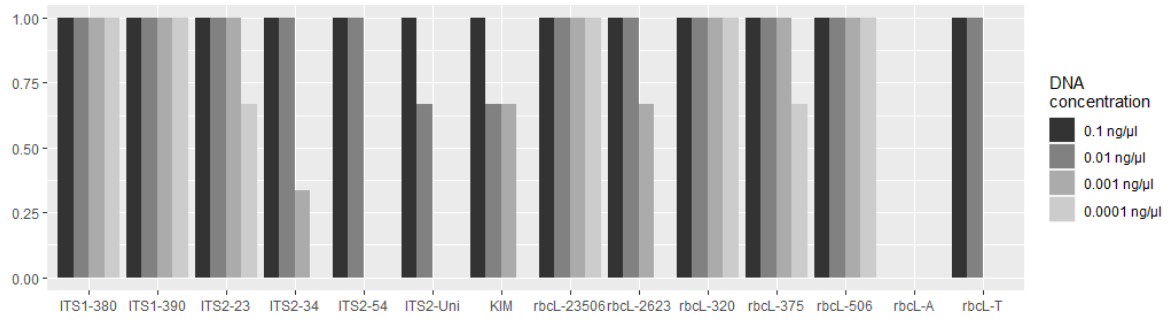

Figure D.3.2.1.610:: Exploratory analysis of amplification success at suboptimal DNA concentrations based on plant DNA extracts from 3 cucurbits (proportion of successful PCRs, n = 3).

The amplification success of each primer was then tested on 6 pollen DNA extracts from *Apis mellifera* collected from cucurbit crops in Tanzania (preserved in EtOH, Qiagen DNA extraction, DNA concentration range = 0.52-0.024 ng/ul). The primer pairs ITS2-23 and rbcL-320 worked with all the 6 pollen DNA extracts tested. The other twelve primer pairs generally worked at a lower success rate (Fig. D.3.2.1.7:).

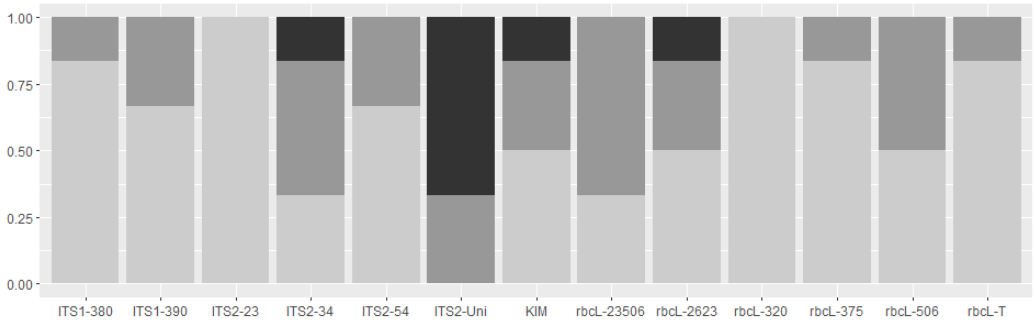

Figure D.3.2.1.711: Proportions of successful (light grey), uncertain (grey) or failed PCRs (black) as obtained using different primers on fresh pollen DNA extracts (n=6).

With the objective of defining cost- and time effective standards for research on pollen and microbial profiling, we preliminarily and qualitatively considered expected performance and costs of common wet lab pipelines for DNA metabarcoding (results not reported here).

We eventually designed two custom wet lab pipelines (Annex 2) complementary or alternative to the popular Nextera XT pipeline (Annex 2) for DNA metabarcoding. The designed pipelines aimed at using different reagents (semi-custom library prep) or different reagents and indexes (fully custom library prep) to be purchased in bulk and used on batches of samples of different sizes (from only a few to a few hundreds). The main rationale of this approach was to achieve a relatively low and uniform cost/sample and to increase scalability compared to a quite expensive commercial kit which only allows processing either 24 or 96 samples.

Due to the relatively high costs of HTS technologies, we adopted a step-by-step approach with the objective of developing the fully custom pipeline only in case the semi-custom protocol would have provided effective advantages in terms of performance or time/cost effectiveness.

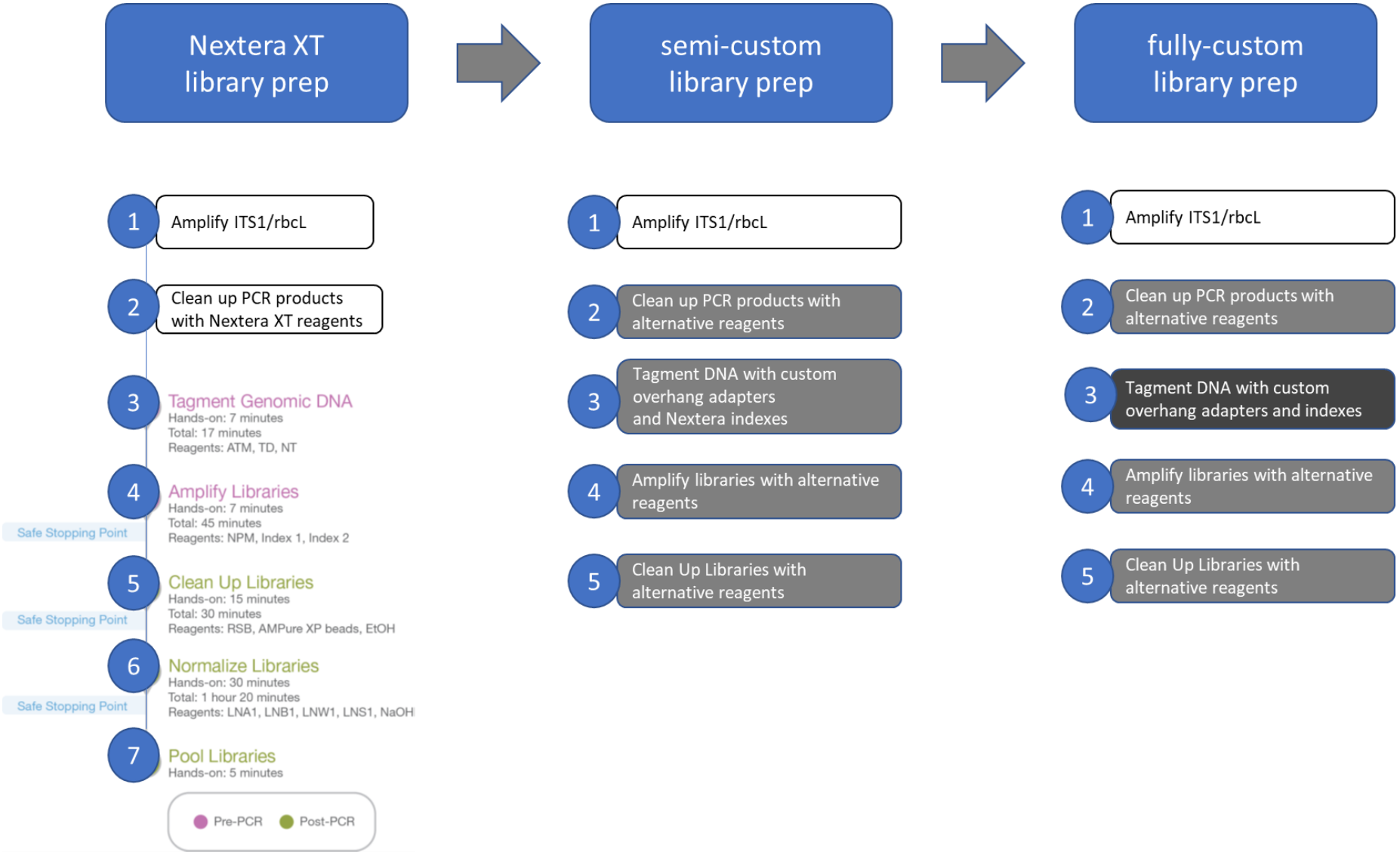

Fig. D.3.2.1.812:: Complementary (semi-custom library prep) or alternative pipelines (fully-custom library prep) to Nextera XT DNA metabarcoding (left)

The semi-custom pipeline included reagents routinely used at RMCA such as the DNA polymerases kit (Invitrogen 10966-050) for 1st amplicon PCR and 2nd indexing PCR (steps 1 and 3 in Fig. D.3.2.1.8.) and the AMPure XP beads (Beckman Coulter A63881) for DNA cleanup and size selection (steps 2 and 5). The fully-custom pipeline also included the use of custom-made dual indexes (including P adaptors) which could have been synthesized and purchased from specialized companies (such as Macrogen or Eurogentec). For a detailed overview of the pipeline (Annex 2).

The performance of the semi-custom and of the Nextera XT library prep were compared during a lab test organized at RMCA in 2022. DNA was extracted from 24 pollen loads from flower flies (Diptera, *Tephritidae*) and bees (Hymenoptera Apoidea) as per manufacturer's instructions of the DNeasy Plant Mini Kit (Qiagen cat. 69106). Following DNA quality check, each DNA extract was aliquoted in 4 samples, which were subjected either to

1. Nextera XT library prep following ITS2 (primerpair ITS2-34, Annex 2) amplification or
2. Nextera XT library prep following rbcL (primerpair rbcL-320, Annex 2) amplification or
3. Semi-custom library prep following ITS2 amplification or

The 96 metagenomic libraries obtained (metadata available in Annex 2) were pooled and, after standardising their DNA concentrations, submitted to Macrogen for High Throughput Sequencing on a single MiSeq flowcell (300 PE, 8Gb output).

The performance of the semi-custom library preparation pipeline was generally lower compared to Nextera XT (Fig. D.3.2.1.9. and D.3.2.1.10.), with 79.2% of semi-custom libraries (n=48) showing lower yields in terms of raw reads, 83.3% in terms of n. of Amplicon Sequence Variants (ASVs) and 52.1% in terms of cumulative n. ASVs. The Nextera XT libraries outperforming semi-custom library prep showed an average gain of raw reads of 29.8% (SD=20.8%), while the average gain of raw and filtered reads in semi-custom library prep libraries outperforming Nextera XT was 8.2% (SD=6.6%). The Nextera XT libraries outperforming semi-custom library prep with respect to the ratio between n. filtered / n. raw reads showed an average gain 44.2% (SD=34.1%), while the average gain of in semi-custom library prep libraries outperforming Nextera XT was 20.7% (SD=17.1%). The average gain in terms of n. of ASVs and cumulative n. of ASVs of Nextera XT libraries outperforming semi-custom library prep was 54.8% (SD=26.6%) and 40.9% (SD=35.8%), while the average gain of in semi-custom library prep libraries outperforming Nextera XT was 20.2% (SD=17.8%) and 37.8% (30.1%).

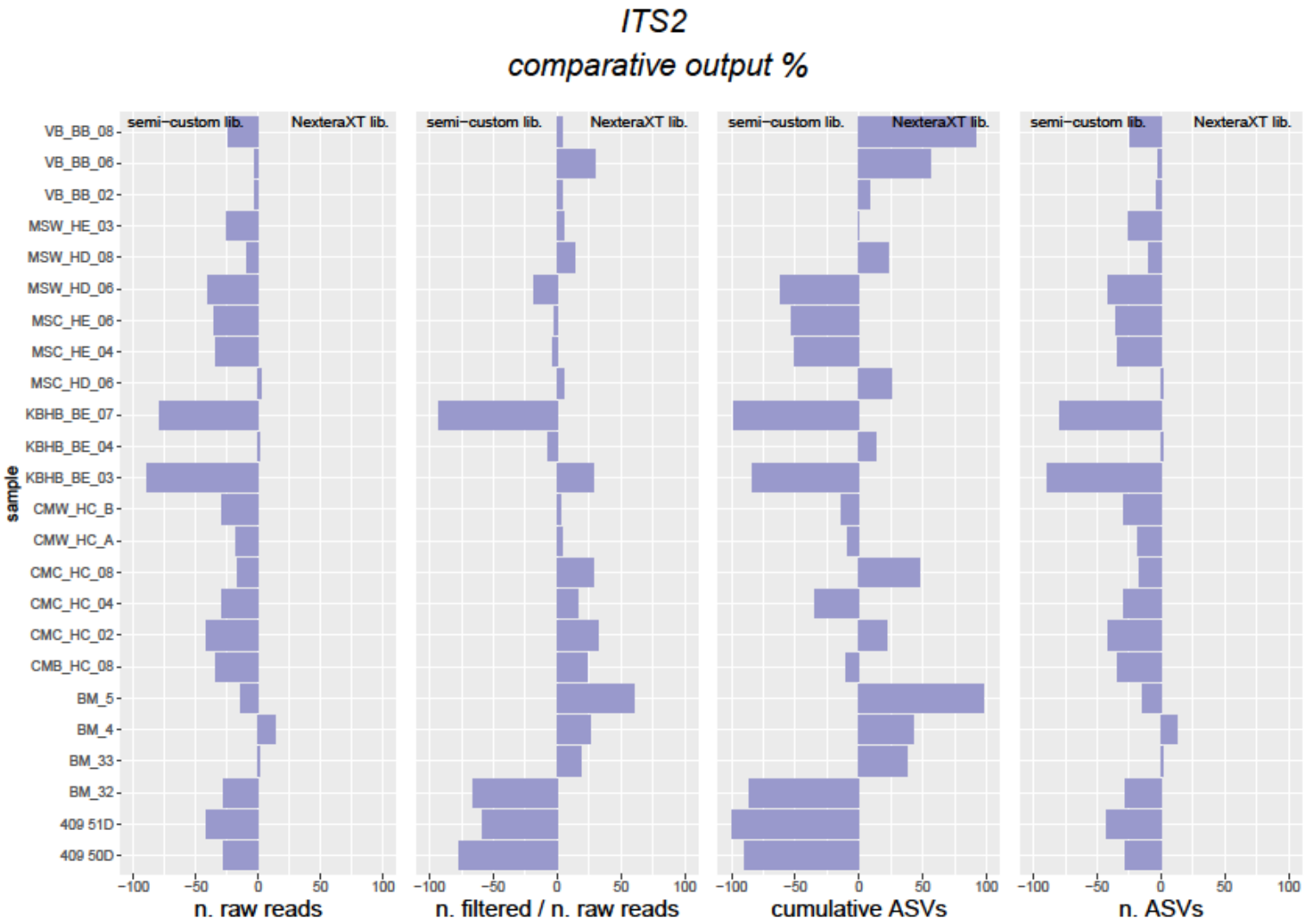

Fig. D.3.2.1.913: Comparative output of DNA extracts subjected to semi-custom or Nextera XT library prep following amplification of ITS2. Gain / loss % in n. filtered / n. reads and in n. of ASVs and cumulative n. of ASVs are shown.

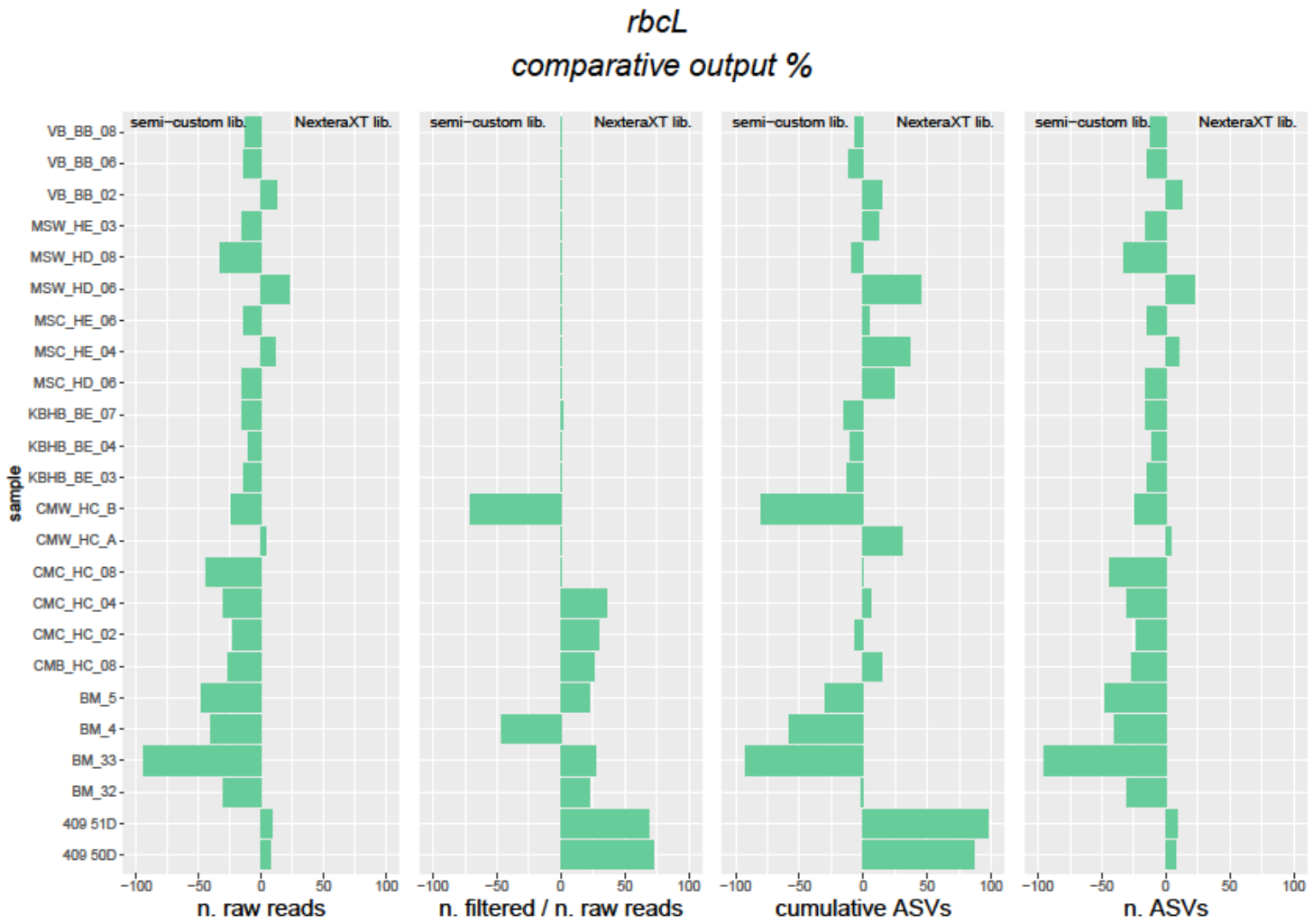

Fig. D.3.2.1.1014.: Comparative output of DNA extracts subjected to semi-custom or Nextera XT library prep following amplification of rbcL. Gain / loss % in n. filtered / n. raw reads and in n. of ASVs and cumulative n. of ASVs are shown.

ANOVA (Tab. D.3.2.1.4:) detected significantly higher number of raw reads in Nextera XT compared to semi-custom library prep, and higher number of ASVs in ITS2 Nextera XT libraries (while no differences were detected between rbcL Nextera XT and semi-custom libraries). Additionally, and irrespectively of library prep, rbcL amplification yielded a higher ratio number filtered / number raw reads and a higher n. of cumulative ASVs compared to ITS amplification.

Tab. D.3.2.1.4: ANOVAs testing the effects of marker amplification (ITS2, rbcl) and library prep (semi-custom, Nextera XT) on (a) n. of raw reads, (b) ratio n. filtered / n. raw reads, (c) n. of ASVs and (d) cumulative n. ASVs in 96 DNA metabarcoding libraries. *: P<0.05, **: P<0.01, ns: not significant. Significant effects are highlighted in yellow.

		df	MS	F	P	
n. raw reads	Marker: MAR	1	0.075	0.025	0.875	ns
	Library prep: LIB	1	38.425	12.711	0.001	*
	MAR:LIB	1	1.780	0.589	0.445	ns
	Residual	92	3.023			
	C = 0.434, p = 0.016, transformation: fourth root					
n. filtered / n. raw reads	Marker: MAR	1	0.892	23.240	0.000	*
	Library prep: LIB	1	0.085	2.221	0.140	ns
	MAR:LIB	1	0.075	1.948	0.166	ns
	Residual	92	0.038			
	C = 0.378, p = 0.114, transformation: none					
n. ASVs	Marker: MAR	1	36.046	19.577	0.000	*
	Library prep: LIB	1	47.077	25.569	0.000	*
	MAR:LIB	1	9.937	5.397	0.022	*
	Residual	92	1.841			
	C = 0.343, p = 0.315, transformation: square root					
	SNK test MAR x LIB:	ITS2	custom lib. < NexteraXT lib.			
		ITS2	custom lib. = NexteraXT lib.			
cumulative n. ASVs	Marker: MAR	1	13824.765	7.427	0.008	*
	Library prep: LIB	1	587.413	0.316	0.576	ns
	MAR:LIB	1	1615.921	0.868	0.354	ns
	Residual	92	1861.516			
	C = 0.336, p = 0.376, transformation: square root					

As the performance of semi-custom library prep was significantly lower than Nextera XT In terms of raw reads and n. of ASVs, and as semi-custom library prep did not prove to be cost or time effective compared to outsourcing (see below), we did not proceed any further with the setting up of a second experiment to test the performance of fully-custom library prep.

Conversely, we proceeded exploring the results obtained from pollen loads from different groups of bees and flower flies. These analyses were implemented on the data obtained using the better performing Nextera XT library prep. In particular, we tested differences between:

- groups of pollen loads (each including 6 replicated libraries) from *Syrphidae* recently collected (2022, fresh flower flies), *Apis mellifera* recently collected (2022, fresh

honeybees) and collection honeybees (Museum honeybees) collected in 1922 (n=2), 1947 (n=2), 1963 (n=1), 1993 (n=1).

- groups of pollen loads (each including 3 replicated libraries) from three species of recently collected *Syrphidae* (*Betasyrphus adligatus* (Wiedemann), *Ischiodon aegyptius* (Wiedemann), *Toxomerus floralis*(Fabricius))

Surprisingly, the first test did not show significant differences between fresh flower flies, fresh honeybees and Museum honeybees (Fig. . D.3.2.1.11. and . D.3.2.1.12., Tab. D.3.2.1.4: and D.3.2.1.5.) either in terms of number of raw reads or of number of ASVs and cumulative ASVs. This was unexpected due to (a) the larger pollen loads which were isolated from fresh honeybees compared to fresh flower flies (as qualitatively observed during pollen isolation) and (b) to the expected lower quality of pollen DNA isolated from honeybees dating back up to 1922. Yet, significant differences were found in the performance of the two markers, with rbcL yielding a higher n. filtered / n. raw reads ratio in fresh bees compared to ITS2.

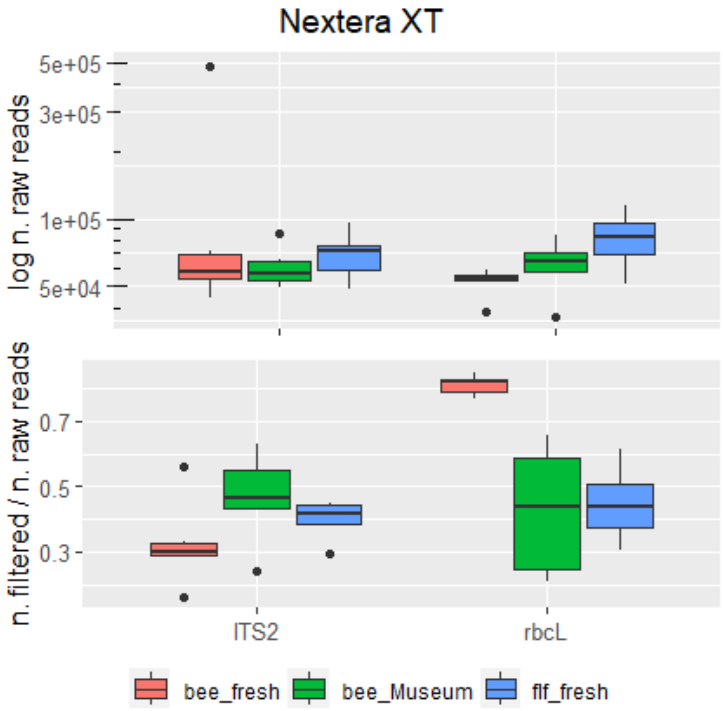

Fig. D.3.2.1.11.: Number of raw and filtered reads obtained from 36 DNA Nextera XT metabarcoding libraries following ITS2 and rbcl amplification in three sample groups including pollen loads from fresh flower flies, fresh honeybees, Museum honeybees.

Tab. D.3.2.1.5: ANOVAs testing the effects of Nextera XT marker amplification (ITS2, rbcl) and sample group (fresh flower flies, fresh honeybees, Museum honeybees) on n. of raw reads and on the ratio n. filtered / n. raw reads in 36 DNA metabarcoding libraries. ns: not significant.

.		df	MS	F	P	
n. raw reads	Marker: MAR	1	1.996	0.480	0.494	ns
	Group: GRO	2	1.961	0.472	0.629	ns
	MAR:GRO	2	6.638	1.596	0.219	ns
	Residual	30	4.158			
	C = 0.811, p = 1.548e-07, transformation: fourth root					
n. filtered / n. raw reads	Marker: MAR	1	0.038	9.450	0.004	**
	Group: GRO	2	0.007	1.782	0.186	
	MAR:GRO	2	0.045	11.374	0.000	***
	Residual	30	0.004			

	C = 0.442, p = 0.051, transformation: fourth root		
	SNK test MAR x GRO:	bee fresh	rbcl > ITS2
		bee Museum	rbcl = ITS2
		flf fresh	rbcl = ITS2

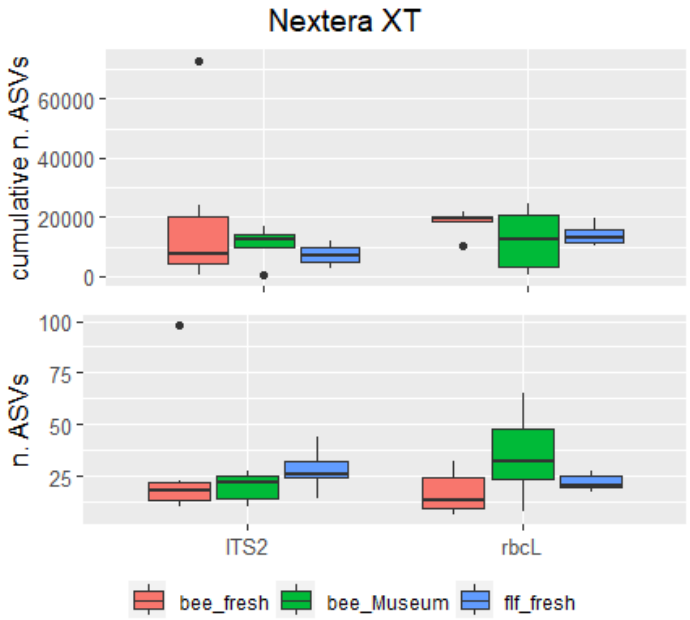

Fig.. D.3.2.1.12.: Number of ASVs and cumulative ASVs obtained from 36 DNA Nextera XT metabarcoding libraries following ITS2 and rbcl amplification in three sample groups including pollen loads from fresh flower flies, fresh honeybees, Museum honeybees.

Tab. D.3.2.1.6: ANOVAs testing the effects of Nextera XT marker amplification (ITS2, rbcl) and sample group (fresh flower flies, fresh honeybees, Museum honeybees) on n. of ASVs and cumulative n. ASVs in 36 DNA metabarcoding libraries. ns: not significant.

		df	MS	F	P	
cumulative n. ASVs	Marker: MAR	1	10.045	1.625	0.212	ns
	Group: GRO	2	5.618	0.909	0.414	ns
	MAR:GRO	2	3.160	0.511	0.605	ns
	Residual	30	6.182			
	C = 0.402, p = 0.11, transformation: fourth root					
n. ASVs	Marker: MAR	1	0.105	0.048	0.828	ns
	Group: GRO	2	1.132	0.517	0.602	ns
	MAR:GRO	2	4.509	2.058	0.145	ns
	Residual	30	2.191			
	C = 0.465, p = 0.03, transformation: square root					

Similarly, we did not observe significant interspecific differences from the DNA metabarcoding of pollen loads from three flower fly species (Fig. D.3.2.1.13. and D.3.2.1.14., Tab. D.3.2.1.6. and D.3.2.1.7.), while as already observed for the main analysis reported in the beginning of this section,

the DNA metabarcoding of rbcL provided significantly higher output in terms of ratio n. filtered / n. raw reads and cumulative n. of ASVs.

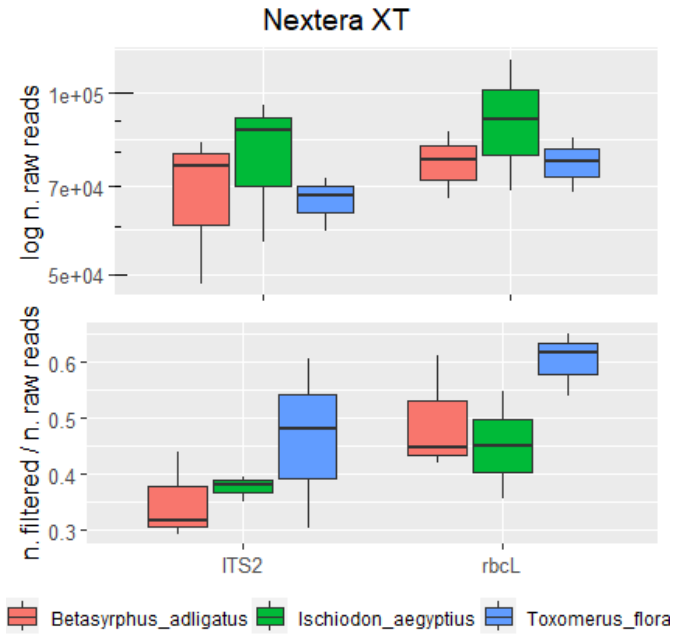

Fig. D.3.2.1.13.: Number of raw reads and on the ratio n. filtered / n. raw reads obtained from 18 DNA Nextera XT metabarcoding libraries following ITS2 and rbcl amplification in three sample groups including pollen loads from recently collected (2022) flower flies from three species (*Betasyrphus adligatus, Ischiodon aegyptius, Toxomerus floralis*).

Tab. D.3.2.1.7: ANOVAs testing the effects of Nextera XT marker amplification (ITS2, rbcl) and flower fly species (*Betasyrphus adligatus, Ischiodon aegyptius, Toxomerus floralis*) on n. of raw reads and on the ratio n. filtered / n. raw reads in 18 DNA metabarcoding libraries. ***: p<0.001, ns: not significant.

		df	MS	F	P	
n. raw reads	MAR	1	4.3E+08	1.776	0.207	ns
	SPE	2	3.5E+08	1.419	0.280	ns
	MAR:SPE	2	4710460	0.019	0.981	ns
	Residual	12	2.4E+08			
	C = 0.353, p = 0.678, transformation: none					
n. filtered / n. raw reads	MAR	1	0.064	7.188	0.020	*
	SPE	2	0.026	2.955	0.090	ns
	MAR:SPE	2	0.002	0.255	0.779	ns
	Residual	12	0.009			
	C = 0.419, p= 0.394, transformation: none					

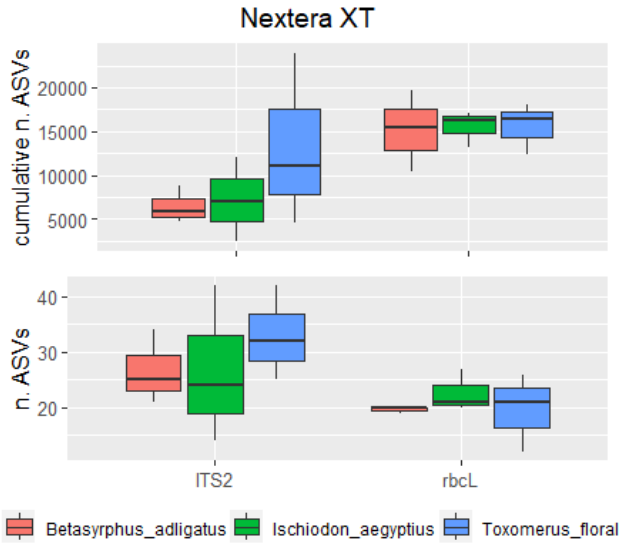

Fig. D.3.2.1.14.: N. of ASVs and cumulative n. of ASVs obtained from 18 DNA Nextera XT metabarcoding libraries following ITS2 and rbcl amplification in three sample groups including pollen loads from recently collected (2022) flower flies from three species (*Betasyrphus adligatus, Ischiodon aegyptius, Toxomerus floralis*).

Tab. 6: ANOVAs testing the effects of Nextera XT marker amplification (ITS2, rbcl) and flower fly species (*Betasyrphus adligatus, Ischiodon aegyptius, Toxomerus floralis*) on n. of ASVs and cumulative n. ASVs in 18 DNA metabarcoding libraries. *: p<0.05, ns: not significant.

		df	MS	F	P	
n. ASVs	MAR	1	296.056	4.62989	0.052	ns
	SPE	2	15.0556	0.23545	0.793	ns
	MAR:SPE	2	34.0556	0.53258	0.600	ns
	Residual	12	63.9444			
	C = 0.526, p = 0.142, transformation: none					

cumulative n. ASVs	MAR	1	1.9E+08	7.174	0.020	*
	SPE	2	2.3E+07	0.854	0.449	ns
	MAR:SPE	2	1.9E+07	0.726	0.503	ns
	Residual	12	2.6E+07			
	C = 0.594, p= 0.065, transformation: none					

These results suggest that DNA metabarcoding could be profitably implemented also from reduced amount of pollen, as is the case for pollen loads isolated from small-sized flower flies or wild bees. Pollen metabarcoding from the insect historical collections of RMCA/RBINS should be technically feasible by using standard and routinely used wet-lab procedures. Its technical feasibility however, would not exclude major problems with sample cross contamination (e.g. across specimens preserved in the same box).

A more in detail analysis on pollen compositional differences across the sample groups targeted by this project as well as on the performance of DNA metabarcoding IDs provided by different markers is currently ongoing in the the framework of the project ISeBAF (project partners M. Virgilio, JEMU RMCA and C. Vangestel, JEMU RBINS). These results will be communicated in the framework of the ongoing collaborative research between RMCA, , the Sokoine University of Agriculture, RBINS and the Botanical garden of Meise.

Besides the technical performance of pipelines for DNA metabarcoding, we also considered the time and cost-effectiveness of outsourcing all or part of the wet-lab pipelines to specialized companies. The main rationale behind this analysis was to reduce as much as possible the working costs for DNA

metabarcoding. The main assumption was that external companies would provide the same (or higher) quality standards and output than those achieved with “in-house” library prep. Working costs were calculated both in terms of lab consumables and of personnel costs for wet-lab time, which were estimated by considering 96 libraries / working week / person, and a gross year salary / person of 60k €.

Below (Fig. D.3.2.1.15.) a schematic representation of cost/sample as calculated for a batch of 96 samples (as this is the sample size commonly loaded on a Miseq lane for HTS) and one marker (e.g. ITS2) is shown. The lowest cost / samples are reported in orange and to an offer for a batch of minimum 227 samples.

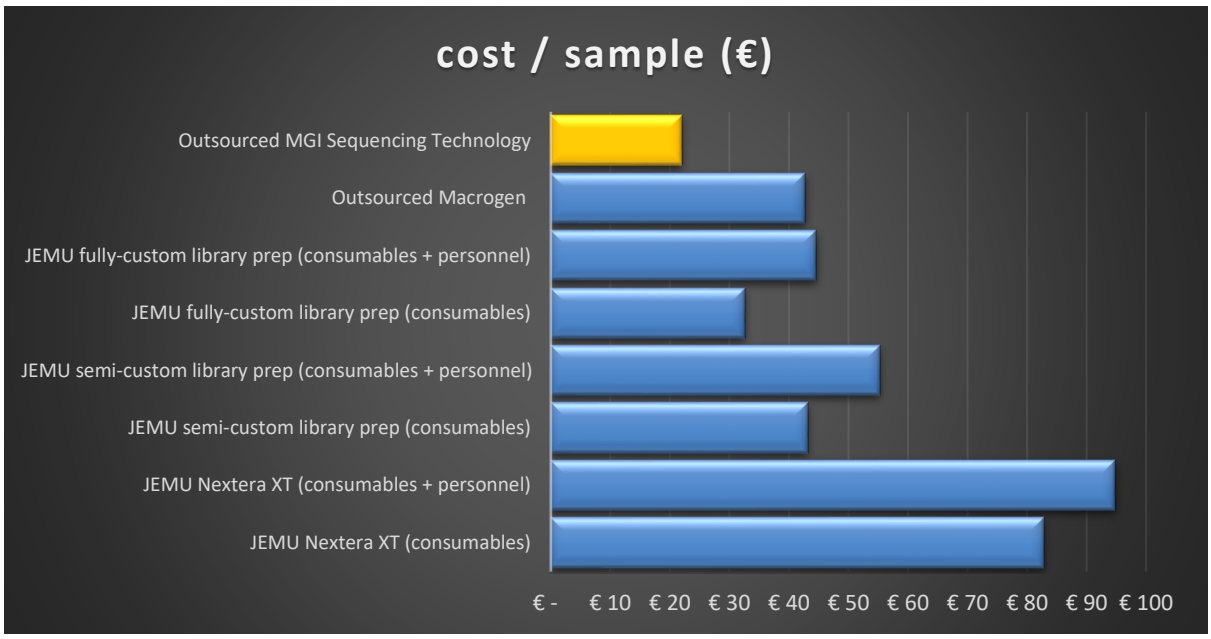

Fig. D.3.2.1.15.: Total cost / sample including library prep and HTS, as calculated for a typical run of 96 samples on a Miseq lane (in blue). The lowest cost/sample is represented in orange and refers to a batch of minimum 227 samples (in yellow). Costs for “in-house” library prep (labelled as “JEMU”) are provided for consumables only and for consumables and personnel costs.

Irrespectively of performance (which was generally lower in the semi-custom pipeline, compared to the pipeline based on a commercial kit and regardless we acknowledge that in-house library prep might still present some advantage in terms of scalability, ad hoc optimization, last minute substitution of a few samples based on QC, etc., it is very clear how outsourcing provides the most convenient option for the DNA metabarcoding of relatively large batches of samples (at least 96 samples, with the lowest costs obtained with an investment of 5k € for 227 samples (Annex 2).

(Also see general recommendations in section 4.2)

Selected References

Bafeel, S. O., Arif, I. A., Bakir, M. A., Khan, H. A., Al Farhan, A. H., Al Homaidan, A. A., Ahamed, A., & Thomas, J. (2011). Comparative evaluation of PCR success with universal primers of maturase K (matK) and ribulose-1, 5-bisphosphate carboxylase oxygenase large subunit (rbcL) for barcoding of some arid plants. Plant OMICS, 4(4), 195–198.

Bell, K. L., De Vere, N., Keller, A., Richardson, R. T., Gous, A., Burgess, K. S., & Brosi, B. J. (2016). Pollen DNA barcoding: Current applications and future prospects. *Genome*, *59*(9), 629–640. https://doi.org/10.1139/gen-2015-0200

Bell, K. L., Loeffler, V. M., & Brosi, B. J. (2017). An rbcL Reference Library to Aid in the Identification of Plant Species Mixtures by DNA Metabarcoding . *Applications in Plant Sciences*, *5*(3), 1600110. https://doi.org/10.3732/apps.1600110

Bruns, T. D., Lee, S. B., & Taylor, J. W. (1990). *Amplification and direct sequencing of fungal ribosomal RNA Genes for phylogenetics*. *May 2014*.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y., & Leon, C. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE*, *5*(1), 1–8. https://doi.org/10.1371/journal.pone.0008613

Cheng, T., Xu, C., Lei, L., Li, C., Zhang, Y., & Zhou, S. (2016). Barcoding the kingdom Plantae: New PCR primers for ITS regions of plants with improved universality and specificity. *Molecular Ecology Resources*, *16*(1), 138–149. https://doi.org/10.1111/1755-0998.12438

Coghlan, S. A., Shafer, A. B. A., & Freeland, J. R. (2021). *Development of an environmental DNA metabarcoding assay for aquatic vascular plant communities*. *June 2020*, 372–387. https://doi.org/10.1002/edn3.120

De Vere, N., Rich, T. C. G., Ford, C. R., Trinder, S. A., Long, C., Moore, C. W., Satterthwaite, D., Davies, H., Allainguillaume, J., Ronca, S., Tatarinova, T., Garbett, H., Walker, K., & Wilkinson, M. J. (2012). *DNA Barcoding the Native Flowering Plants and Conifers of Wales*. *7*(6), 1–12. https://doi.org/10.1371/journal.pone.0037945

De Vere, N., Jones, L. E., Gilmore, T., Moscrop, J., Lowe, A., Smith, D., Hegarty, M. J., Creer, S., & Ford, C. R. (2017). Using DNA metabarcoding to investigate honey bee foraging reveals limited flower use despite high floral availability. *Scientific Reports*, *7*(January), 1–10. https://doi.org/10.1038/srep42838

Doyle, J. J., & Doyle, J. L. (1987). Doyle_plantDNAextractCTAB_1987.pdf. In *Phytochemical Bulletin* (Vol. 19, Issue 1, pp. 11–15). https://webpages.uncc.edu/~jweller2/pages/BINF8350f2011/BINF8350_Readings/Doyle_plantDNAextractCTAB_1987.pdf

Erickson, D. L., Reed, E., Ramachandran, P., Bourg, N. A., McShea, W. J., & Ottesen, A. (2017). Reconstructing a herbivore's diet using a novel rbcL DNA mini-barcode for plants. *AoB PLANTS*, *9*(3). https://doi.org/10.1093/aobpla/plx015

Fazekas, A. J., Burgess, K. S., Kesanakurti, P. R., Graham, S. W., Newmaster, S. G., Husband, B. C., Percy, D. M., Hajibabaei, M., & Barrett, S. C. H. (2008). Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE*, *3*(7). https://doi.org/10.1371/journal.pone.0002802

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, *3*(5), 294–299. https://doi.org/10.1071/ZO9660275

García-Robledo, C., Erickson, D. L., Staines, C. L., Erwin, T. L., & Kress, W. J. (2013). Tropical Plant-Herbivore Networks: Reconstructing Species Interactions Using DNA Barcodes. *PLoS ONE*, *8*(1). https://doi.org/10.1371/journal.pone.0052967

Gous, A., Swanevelder, D. Z. H., Eardley, C. D., & Willows-Munro, S. (2019). Plant–pollinator interactions over time: Pollen metabarcoding from bees in a historic collection. *Evolutionary Applications*, *12*(2), 187–197. https://doi.org/10.1111/eva.12707

Henry, R. J. (2009). Plant DNA extraction. *Plant Genotyping: The DNA Fingerprinting of Plants*, 239–249. https://doi.org/10.1079/9780851995151.0239

Kress, W. J., & Erickson, D. L. (2007). A Two-Locus Global DNA Barcode for Land Plants: The Coding rbcL Gene Complements the Non-Coding trnH-psbA Spacer Region. *PLoS ONE*, *2*(6). https://doi.org/10.1371/journal.pone.0000508

Laha, R. C., Mandal, S. De, Ralte, L., Ralte, L., & Kumar, N. S. (2017). Meta - barcoding in combination with palynological inference is a potent diagnostic marker for honey floral composition. *AMB Express*. https://doi.org/10.1186/s13568-017-0429-7

Lalhmangaihi, R., Ghatak, S., Laha, R., Gurusubramanian, G., & Senthil Kumar, N. (2014). Protocol for optimal quality and quantity pollen DNA isolation from Honey samples. *Journal of Biomolecular Techniques*, *25*(4), 92–95. https://doi.org/10.7171/jbt.14-2504-001

Lucas, A., Bodger, O., Brosi, B. J., Ford, C. R., Forman, D. W., Greig, C., Hegarty, M., Neyland, P. J., & de Vere, N. (2018). Generalisation and specialisation in hoverfly (*Syrphidae*) grassland pollen transport networks revealed by DNA metabarcoding. *Journal of Animal Ecology*, *87*(4), 1008–1021. https://doi.org/10.1111/1365-2656.12828

Moorhouse-Gann, R. J., Dunn, J. C., Vere, N. de, Goder, M., Cole, N., Hipperson, H., & Symondson, W. O. C. (2018). New universal ITS2 primers for high-resolution herbivory analyses using DNA metabarcoding in both tropical and temperate zones. *Scientific Reports*, *November 2017*, 1–15. https://doi.org/10.1038/s41598-018-26648-2

Newmaster, S. G., Fazekas, A. J., & Ragupathy, S. (2006). DNA barcoding in land plants: Evaluation of rbcL in a multigene tiered approach. *Canadian Journal of Botany*, *84*(3), 335–341. https://doi.org/10.1139/B06-047

Peel, N., Dicks, L. V., Clark, M. D., Heavens, D., Percival-Alwyn, L., Cooper, C., Davies, R. G., Leggett, R. M., & Yu, D. W. (2018). *Semi-quantitative characterisation of mixed pollen samples using MinION sequencing and Reverse Metagenomics*. 2–41.

Prosser, S. W. J., & Hebert, P. D. N. (2017). Rapid identification of the botanical and entomological sources of honey using DNA metabarcoding. *Food Chemistry*, *214*, 183–191. https://doi.org/https://doi.org/10.1016/j.foodchem.2016.07.077

Richardson, R. T., Lin, C.-H., Sponsler, D. B., Quijia, J. O., Goodell, K., & Johnson, R. M. (2015). Application of ITS2 Metabarcoding to Determine the Provenance of Pollen Collected by Honey Bees in an Agroecosystem. *Applications in Plant Sciences*, *3*(1), 1400066. https://doi.org/10.3732/apps.1400066

Sandrini-Neto, L., & Camargo, M. G. (2015). GAD: an R package for ANOVA designs from general principles. *Available on CRAN*Vere, N. De, Rich, T. C. G., Ford, C. R., Trinder, S. A., Long, C., Moore, C. W., Satterthwaite, D., Davies, H., Allainguillaume, J., Ronca, S., Tatarinova, T., Garbett, H., Walker, K., & Wilkinson, M. J. (2012). *DNA Barcoding the Native Flowering Plants and Conifers of Wales*. *7*(6), 1–12. https://doi.org/10.1371/journal.pone.0037945

White, T. J., Bruns, T. D., Lee, S. B., & Taylor, J. W. (1990). *Amplification and direct sequencing of fungal ribosomal RNA Genes for phylogenetics*.

Wilson, E. E., Sidhu, C. S., Levan, K. E., & Holway, D. A. (2010). Pollen foraging behaviour of solitary Hawaiian bees revealed through molecular pollen analysis. *Molecular Ecology*, *19*(21), 4823–4829. https://doi.org/10.1111/j.1365-294X.2010.04849.x

Zhang, G., Liu, J., Gao, M., Kong, W., Zhao, Q., Shi, L., & Wang, Q. (2020). Tracing the Edible and Medicinal Plant Pueraria montana and Its Products in the Marketplace Yields Subspecies Level Distinction Using DNA Barcoding and DNA Metabarcoding. In *Frontiers in Pharmacology* (Vol. 11). https://www.frontiersin.org/article/10.3389/fphar.2020.00336

D 3.3.1: production of digital vouchers complementing the RMCA collections (M24)

Digital images from > 700 insect vouchers were produced and linked to the collection, genomic, and DNA vouchers. The list of digital images produced is provided as Annex 3. A subset of images from > 550 *Dacus* and *Ceratitis* (Diptera, *Tephritidae*) vouchers was also selected for publication on the https://fruitflies.africamuseum.be/ website which provides links to the voucher metadata and digital images, as well as to the Darwin interface.

D 3.3.2: production of DNA collection vouchers complementing the RMCA collections (M24).

DNA vouchers from > 700 insect vouchers were produced and linked to the collection, genomic, and digital vouchers. The list of DNA vouchers produced is provided as Annex 3. The DNA collections of RMCA are maintained through dedicated long-term stabilization protocols (https://gentegra.com/gentegra-dna-2/).

D 3.3.3: establishing a collection of RMCA genomic vouchers (M24)

Genomic vouchers from > 1300 insect vouchers from the RMCA insect collections were produced and linked to the collection, DNA, and digital vouchers. The list of genomic vouchers produced is provided as Annex 3.

D 3.3.4: online, open access database of genomic vouchers (M24)

The list of genomic vouchers, including WGS data and the voucher metadata (these latter linked to the corresponding insect and DNA vouchers is available as a downloadable database (see D.5.1.1). A link to the published, open-access WGS data will be gradually made available following publication of results from the different research projects listed in Section 7 Acknowledgements. All genomic vouchers are available for collaborative research upon request to the project coordinators.

WP 4: Coordination, project management and reporting

Task 4.1: workplan implementation

The reference taxonomists for syrphids and tephritids coordinated and supervised the selection of Museum vouchers processed and provided valuable input to the setup of the experimental tests. The JEMU of RMCA/RBINS, addressed all aspects related to the setup of the experimental design, to their practical implementation, to the analysis of the experimental data and to the WGS and genome assemblage of the targeted vouchers. The two responsible persons of the molecular labs at RMCA and RBINS contributed to the supervision and coordination of the lab activities and, as part of their institutional tasks, to the routine functioning of the laboratories. The JEMU prepared and submitted the periodical project reports and organised the meetings for the follow-up committee in which the suitability of methods and approaches were discussed. The ICT service of RMCA, in collaboration with the JEMU-RMCA, provided assistance in the upload of the open access database on the RMCA website.

D 4.1.1: meetings follow-up committee (M1, M12, M24)

Report of the 1st meeting of the project follow-up Committee (July 2021) available as Annex 4, Annex 5.

Report of the 2nd meeting of the project follow-up Committee (Dec 2022) available as Annex 6, Annex 7.

Report of the 3rd meeting of the project follow-up Committee (Jul 2023) available as Annex 8, Annex 9.

D 4.1.2: project reports (initial, annual, final) (M3, M12, M24)

Initial report available as Annex 10.

Annual report 2022 available as Annex 11.

Final report 2023: present document.

WP 5: Data management

Task 5.1: data and metadata generated for digital and genomic vouchers

The WGS data generated were linked to the corresponding metadata and incorporated into the RMCA collections as genomic vouchers. Similarly, digital images were associated to the corresponding image metadata and incorporated into the RMCA collections of digital vouchers. All metadata from the morphological, digital and genomic vouchers were linked to the corresponding collection specimen accession numbers and included in the open access database available on the RMCA website.

D 5.1.1: backup and long-term storage of WGS data and metadata (M24)

The backup and storage of WGS data in the Network-attached storage (NAS) systems of RMCA has been finalised for > 1300 vouchers from the RMCA insect collection. List of samples, metadata and pathways to the NAS files are provided as Annex 3. These data were also formatted according to the standards required for publication on the DARWIN interface of RMCA/RBINS. DARWIN input file available as Annex 12.

D 5.1.2: backup and long-term storage of digital images and metadata (M24)

The backup and storage of digital images from > 700 insect vouchers in the NAS systems of RMCA has been finalised. List of files, metadata and pathways to the NAS files are provided as Annex 3. These data were also formatted according to the standards required for publication on the DARWIN interface of RMCA/RBINS. DARWIN input file available as Annex 12. A subset of images from > 550 *Dacus* and *Ceratitis* (Diptera, *Tephritidae*) vouchers was also selected for publication on the https://fruitflies.africamuseum.be/ website which provides links to the voucher metadata and digital images, as well as to the Darwin interface.

WP 6: Valorisation, dissemination, exploitation of results

Task 6.1: actions targeting the scientific community

The results of the experimental tests were published in abstracts submitted to an international scientific conferences (11th ISFFEI) and on a poster for an institutional meeting (RMCA Science Days). A methodological paper on Museomics (minimum goal) was prepared, submitted and published on an international scientific journal with IF. A second paper is in preparation.

Task 6.2: actions targeting the general public

Awareness about the importance of preserving and valorising the genetic resources associated to the biological collection of RMCA was raised by posting newsflashes on the RMCA website and/or on social medias.

D 6.1.1: participation to international congresses (M10, M19)

Due travel restrictions related to the COVID-19 emergency, we have not taken part to the international congresses in 2021. In November 2022, the JEMU-RMCA (InsectMOoD coordinator) participated to the 11th International Symposium on Fruit Flies of Economic Importance (ISFFEI, 13-8/11/2022) in Sydney, Australia, Macquarie University. There they presented an abstract directly related to the project and several others, dealing with the genomic vouchers archived by InsectMOod (see section 6).

D 6.1.2: post / interviews on RMCA social media accounts (M6, M12, M18, M24)
Twitter:

Info and newsflashes on RMCA / RBINS websites:

- https://www.africamuseum.be/en/staff/896/project_detail_view?prjid=722
- https://fruitflies.africamuseum.be/activities/projects
- 8 Nov 2022: https://fruitflies.africamuseum.be/news/2022-11-10-isffei

- 9 Nov 2022: https://fruitflies.africamuseum.be/news/2022-11-09-insectmood
- https://jemu.myspecies.info/sites/jemu.myspecies.info/files/Science_days_2022_Poster_InsectMOoD.pdf
- https://jemu.myspecies.info/sites/jemu.myspecies.info/files/Esselens_Poster%2011th%20ISFFEI%20-%20Sydney%20LinkingGenomesToMuseumVouchers.pdf

Info and newsflashes on social media:

Twitter

- 18 Mar 2022: https://twitter.com/EsselensLore/status/1504795410061303813
- 30 Apr 2022: https://twitter.com/EsselensLore/status/1520375177703526403?s=20
- 1 Jul 2022: https://x.com/EsselensLore/status/1542834513448996866?s=20
- 30 Oct 2023: https://twitter.com/EsselensLore/status/1718919763861143881?s=20
- 6 Nov 2023: https://twitter.com/BioDataJournal/status/1721483107444859091?s=20

Linkedin

- 30 Oct 2023: https://www.linkedin.com/feed/update/urn:li:activity:7124688688801853442/

Info available on Institutional websites

BELSPO:

- https://www.belspo.be/belspo/fedra/proj.asp?l=en&COD=B2%2F202%2FP2%2FInsectMOoD

AfricaMuseum:

- https://www.africamuseum.be/en/staff/896/project_detail_view?prjid=722
- https://fruitflies.africamuseum.be/activities/projects

4. RECOMMENDATIONS

The costs directly related to genomic library preparation and sequencing represent one of the main limiting factors hampering the high throughput sequencing (HTS) of large numbers of museum specimens. Until recently, the partial sequencing of genomes, via approaches, such as reduced representation libraries (Ewart et al. 2019) or mitochondrial genomics (Timmermans et al. 2016), was considered as the only suitable approach to build up relatively large genomic datasets. However, the rapid technological advances over the past few years have led to a substantial reduction in costs, so that the large-scale whole genome sequencing (WGS) of vouchers represents a realistic perspective for the valorisation of museum collections (e.g. Crampton-Platt et al. (2016), Malakasi et al. (2019), Strijk et al. (2020)). InsectMOoD provides encouraging indications about the routine collection of genomic data from insect museum vouchers, where with "routine genotyping" we refer to standard and commonly- used wet-lab pipelines for WGS rather than to more elaborate, expensive and technically-challenging protocols which are used to recover highly-degraded DNA (reviewed in Orlando et al. (2021)). We believe that theroutine production of genomic-vouchers would provide a remarkable added value to the insect collections of RMCA and to the museum collections in general. In fact, the approach implemented in InsectMOoD allowed delivering in a relatively short time and with relatively limited resources, a large bulk of easily accessible and properly archived genomic data that are now available for national and international research collaborations. The optimization of the experimental protocols and the collection of the genomic data from Syrphidae and Tephritidae, coordinated by the Joint Experimental Molecular Unit (JEMU) of RMCA and RBINS, further strengthen the expertise of the JEMU in Museomics and generated guidelines of general interest for the high throughput sequencing of biological collections of RMCA and RBINS.

Target taxa: Tephritidae, Syrphidae

The results of this project indicate that standard DNA extraction, based on commercially available kits followed by WGS at 10x genome coverage, represents a cost/time-effective, pragmatic approach to the routine, large-scale genotyping of the Tephritidae and Syrphidae collected over the past few decades (see decision map D.2.2.2). The DNA of diverse and heterogeneously collected samples from the Tephritidae and Syrphidae collections of RMCA, even if generally suboptimal in terms of concentration, fragmentation and contamination, can still generate substantial amounts of quality HTS reads which are suitable for genomic research. Based on the results detailed on WP2, the main recommendations for the WGS of Tephritidae and Syrphidae from the RMCA collections are:

- The standard and commercially available DNeasy Blood and Tissue Kit (Qiagen) provides a cost-effective method of extracting DNA from collection specimens sampled over the past three decades (see D.2.1.2.1),
- Suboptimal samples, although containing fragmented DNA, represent a tractable tissue source for large-scale HTS projects based on standard genomic library preparation (see D.2.1.2.2),
- Outsourcing genomic library preparation and WGS of large batches of samples to external companies seems the most time- and cost-effective approach.
- We recommend a two-step approach, including (a) the use of commercial kits and standard genomic library preparation protocols for a first, general screening of subsets of vouchers and (b) more specialised protocols (also including aDNA methodologies) to be used only for the more problematic specimens which did not yield satisfactory results with routine methodologies.

We believe that this approach represents a pragmatic and cost-effective route to the large-scale genotyping of our insect collections.

Non-target taxa

(a) Non-insect collection vouchers (*Tyto alba*, Tytonidae, Aves)

- DdRAD can not be routinely applied on large museum collections to obtain population-level genomic data.
- Yet, it remains feasible when considering stringent sample selection criteria. Fragmentation assessments of museum samples are highly advised to be implemented in wet lab protocols prior to ddRAD sequencing.
- Such screening is relatively easy to accomplish at minimal cost by any moderately equipped molecular lab and will substantially reduce the risk of both data loss and unnecessary library preparation and sequencing costs.
- Include recent high quality samples as a 'reference' to aid targeting endogenous sequence data in museum specimens.

(b) Pollen recovered from insect collection vouchers

Based on the results obtained, EtOH sample preservation, followed by Qiagen DNeasy Plant Mini Kit DNA extraction seems to be the most suitable combination for pollen DNA barcoding in flower flies.

Conversely, CTAB preservation and CTAB DNA extraction provides inconsistent results, due to cross contamination between the insect voucher and the pollen recovered from its body.

Furthermore, compared to CTAB, the Qiagen protocol is faster (2-3 hours vs 4-6 hours), highly standardised and safer for the health of the operator (as not using β-mercaptoethanol). But in comparison, far more expensive.

The comparisons implemented between "in-house" metabarcoding library prep using a popular commercial kit and a semi-custom made pipeline showed that this latter seems to have poorer performances compared to library prep based on commercial kit. However, even if we observed a significantly lower output in n. of raw reads and cumulative n. of ASVs, both the semi-custom pipeline and the commercial kit have comparable output in terms of n. of ASVs recovered and of ratio n. filtered / n. raw reads. Regardless of these differences, outsourcing seems to provide the most cost- and time-effective approach to DNA metabarcoding, and should be currently considered as the best option particularly for the routine processing of large batches of samples.

5. DISSEMINATION AND VALORISATION

See info detailed in WP6 and in the following section.

6. PUBLICATIONS

On scientific Journals with Impact Factor

(see Annex 13)

Ferrari G., Esselens L., Hart M. L., Janssens S., Kidner C., Mascarello M., Peñalba J. V, Pezzini F., Von Rintelen T., Sonet G., Vangestel C., Virgilio M., Hollingsworth P. M. (2023). Developing the protocol infrastructure for DNA sequencing natural history collections. *Biodiversity Data Journal*, *11*, e102317-. https://doi.org/10.3897/BDJ.11.E102317. IF = 1.3

In preparation

Esselens L., Sonet G., Addison P., Bakengesa J., Bota L., Canhanga L., Cugala D., Daniel B., Delatte H., Deschepper P., Karsten M., Kudra A., Majuba R., Manrakhan A., Muzumbe M., Mwatawala M., Terblanche J., Vanbergen S., Backeljau, T., Jordaens K., De Meyer M., Virgilio, M., Vangestel, C. Linking genomes to Museum collection vouchers.

Abstracts in International Scientific Congresses

(see Annex 14)

Esselens L., Ody J., Sonet G., Deschepper P., Vanbergen S., Ferrari G., Hollingsworth P., De Meyer M., Vangestel C., Virgilio M. (2022). Linking genomes to Museum vouchers: an open-access database for African "true" fruit flies (Diptera, *Tephritidae*). 11th International Symposium on Fruit Flies of Economic Importance (ISFFEI). Sydney, Australia 13-18 November 2022.

Esselens L., Mullens N., Addison P., Canhanga L., Cugala D., Deschepper, P., Karsten M., Manrakhan A., Snyman M., Tsatsu S., Terblanche J.S., Vanbergen S., De Meyer M., Virgilio M. (2022) Genomic characterisation of the population structure of *Ceratitis rosa* and *C. quilicii* (Diptera, *Tephritidae*) in southern Africa. 11th International Symposium on Fruit Flies of Economic Importance (ISFFEI). Sydney, Australia 13-18 November 2022.

Mullens N., Esselens L., White I., Mwatawala M., Svardal H., Virgilio M., De Meyer M. (2022). A phylogenomic approach to better understand the evolution of host-plant preferences in frugivorous *Dacus* (*Tephritidae*). 11th International Symposium on Fruit Flies of Economic Importance (ISFFEI). Sydney, Australia 13-18 November 2022.

7. ACKNOWLEDGEMENTS

We would like to thank the InsectMOoD follow-up committee, **Katerina Guschanski** (Uppsala University, Sweden), **Peter Hollingsworth** (Royal Botanic Garden Edinburgh,UK), **Francesca Scolari** (National Research Council, Pavia, Italy) for sharing their valuable expertise and for critical and positive input to this project.

This work benefited of synergies and co-financing from the projects DISPEST, AGROVEG and DIPODIP (framework agreement 2019-2023, Royal Museum for Central Africa - Directorate-general Development Cooperation), FFI-PM (EU, H2020, grant 818184), REACT (EU, H2020, grant 101059523) and SYNTHESYS+ (EU, H2020, grant 823827).

The intellectual and physical property of samples used in this study are regulated by the Nagoya Protocol on Access and Benefit-sharing or, when this was not possible, by Mutually Agreed Terms (MATs) on the use of genetic resources which are inspired by the principles of the Nagoya Protocol. All documents on intellectual and physical property of samples are available upon request to the authors.

ANNEXES

1. Annex 1_abstract EN_FR_N_executive summary.pdf
2. Annex 2_POLBEN.pdf
3. Annex 3_NAS_data.xlsx
4. Annex 4_report_1st meeting follow up committee.pdf
5. Annex 5_Presentation_1st meeting follow up committee.pdf
6. Annex 6_InsectMOoD_report 2022 follow up committee.pdf
7. Annex_7_Presentation_follow-up_committee_Dec22.pdf
8. Annex 8_InsectMOoD_report 2023 follow up committee.pdf
9. Annex 9_Presentation_follow up_committee 2023.pdf

10. Annex 10_InitialReport_F_N_RMCA.pdf
11. Annex 11_INsectMOoD_AnnualNetworkReport_2021.pdf
12. Annex 12_Darwin_data.xlsx
13. Annex 13_BDJ_article_102317.pdf
14. Annex 14_11ISFFEI-Abstract-Book.pdf

REFERENCES

Van Belleghem, S. M. *et al.* (2018) 'Evolution at two time frames: Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution', *PLOS Genetics*, 14(11), p. e1007796. doi: 10.1371/JOURNAL.PGEN.1007796.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/BIOINFORMATICS/BTU170.

Card, D. C. *et al.* (2021) 'Museum Genomics', *https://doi.org/10.1146/annurev-genet-071719-020506*, 55, pp. 633–659. doi: 10.1146/ANNUREV-GENET-071719-020506.

Catchen, J. M. *et al.* (2011) 'Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences', *G3 Genes|Genomes|Genetics*, 1(3), pp. 171–182. doi: 10.1534/G3.111.000240.

Chen, S. *et al.* (2010) 'Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species', *PLOS ONE*, 5(1), p. e8613. doi: 10.1371/JOURNAL.PONE.0008613.

Colella, J. P., Tigano, A. and MacManes, M. D. (2020) 'A linked-read approach to museomics: Higher quality de novo genome assemblies from degraded tissues', *Molecular Ecology Resources*, 20(4), pp. 856–870. doi: 10.1111/1755-0998.13155.

Crampton-Platt, A. *et al.* (2016) 'Mitochondrial metagenomics: letting the genes out of the bottle', *GigaScience*, 5(1). doi: 10.1186/S13742-016-0120-Y.

Danecek, P. *et al.* (2011) 'The variant call format and VCFtools', *Bioinformatics*, 27(15), p. 2156. doi: 10.1093/BIOINFORMATICS/BTR330.

Davey, J. L. and Blaxter, M. W. (2010) 'RADSeq: next-generation population genetics', *Briefings in Functional Genomics*, 9(5–6), p. 416. doi: 10.1093/BFGP/ELQ031.

Ewart, K. M. *et al.* (2019) 'Museum specimens provide reliable SNP data for population genomic analysis of a widely distributed but threatened cockatoo species', *Molecular Ecology Resources*, 19(6), pp. 1578–1592. doi: 10.1111/1755-0998.13082.

Ferrari, G. *et al.* (2023) 'Developing the Protocol Infrastructure for DNA Sequencing Natural History Collections', *Biodiversity Data Journal 11: e102317*, 11, pp. e102317-. doi: 10.3897/BDJ.11.E102317.

Folmer, O. *et al.* (1994) 'DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates.', *Molecular marine biology and biotechnology*, 3(5), pp. 294–299. doi: 10.1071/ZO9660275.

García-Robledo, C. *et al.* (2013) 'Tropical Plant-Herbivore Networks: Reconstructing Species Interactions Using DNA Barcodes', *PLoS ONE*, 8(1). doi: 10.1371/journal.pone.0052967.

Gous, A. *et al.* (2019) 'Plant–pollinator interactions over time: Pollen metabarcoding from bees in a historic collection', *Evolutionary Applications*, 12(2), pp. 187–197. doi: 10.1111/eva.12707.

Graham, C. F. *et al.* (2015) 'Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq)', *Molecular ecology resources*, 15(6), pp. 1304–1315. doi: 10.1111/1755-0998.12404.

Guschanski, K. *et al.* (2013) 'Next-Generation Museomics Disentangles One of the Largest Primate

Radiations', *Systematic Biology*, 62(4), pp. 539–554. doi: 10.1093/SYSBIO/SYT018.

Holmes, M. W. *et al.* (2016) 'Natural history collections as windows on evolutionary processes', *Molecular ecology*, 25(4), pp. 864–881. doi: 10.1111/MEC.13529.

Knyshov, A., Gordon, E. R. L. and Weirauch, C. (2019) 'Cost-efficient high throughput capture of museum arthropod specimen DNA using PCR -generated baits', *Methods in Ecology and Evolution*. Edited by A. Baselga, 10(6), pp. 841–852. doi: 10.1111/2041-210X.13169.

Knyshov, A., Hoey-Chamberlain, R. and Weirauch, C. (2019) 'Hybrid enrichment of poorly preserved museum specimens refines homology hypotheses in a group of minute litter bugs (Hemiptera: Dipsocoromorpha: Schizopteridae)', *Systematic Entomology*, 44(4), pp. 985–995. doi: 10.1111/syen.12368.

Kress, W. J. and Erickson, D. L. (2007) 'A Two-Locus Global DNA Barcode for Land Plants: The Coding rbcL Gene Complements the Non-Coding trnH-psbA Spacer Region', *PLoS ONE*, 2(6). doi: 10.1371/journal.pone.0000508.

Laha, R. C. *et al.* (2017) 'Meta-barcoding in combination with palynological inference is a potent diagnostic marker for honey floral composition', *AMB Express*, 7(1), pp. 1–8. doi: 10.1186/S13568-017-0429-7/FIGURES/2.

Lang, P. L. M. *et al.* (2020) 'Hybridization ddRAD-sequencing for population genomics of nonmodel plants using highly degraded historical specimen DNA', *Molecular Ecology Resources*, 20(5), pp. 1228–1247. doi: 10.1111/1755-0998.13168.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), p. 2078. doi: 10.1093/BIOINFORMATICS/BTP352.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics (Oxford, England)*, 25(14), pp. 1754–1760. doi: 10.1093/BIOINFORMATICS/BTP324.

Lucena-Aguilar, G. *et al.* (2016) 'DNA Source Selection for Downstream Applications Based on DNA Quality Indicators Analysis', *Biopreservation and biobanking*, 14(4), pp. 264–270. doi: 10.1089/BIO.2015.0064.

Malakasi, P. *et al.* (2019) 'Museomics Clarifies the Classification of Aloidendron (Asphodelaceae), the Iconic African Tree Aloes', *Frontiers in Plant Science*, 10, p. 1227. doi: 10.3389/fpls.2019.01227.

McKenna, A. *et al.* (2010) 'The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data', *Genome research*, 20(9), pp. 1297–1303. doi: 10.1101/GR.107524.110.

Moorhouse-Gann, R. J. *et al.* (2018) 'New universal ITS2 primers for high-resolution herbivory analyses using DNA metabarcoding in both tropical and temperate zones', *Scientific Reports 2018 8:1*, 8(1), pp. 1–15. doi: 10.1038/s41598-018-26648-2.

Nadeau, N. J. *et al.* (2014) 'Population genomics of parallel hybrid zones in the mimetic butterflies, H. melpomene and H. erato', *Genome research*, 24(8), pp. 1316–1333. doi: 10.1101/GR.169292.113.

Newmaster, S. G., Fazekas, A. J. and Ragupathy, S. (2006) 'DNA barcoding in land plants: Evaluation of rbcL in a multigene tiered approach', *Canadian Journal of Botany*, 84(3), pp. 335–341. doi: 10.1139/B06-047.

Peterson, B. K. *et al.* (2012) 'Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species', *PLOS ONE*, 7(5), p. e37135. doi: 10.1371/JOURNAL.PONE.0037135.

Puritz, J. B. *et al.* (2014) 'Demystifying the RAD fad', *Molecular ecology*, 23(24), pp. 5937–5942. doi: 10.1111/MEC.12965.

Sandrini-Neto, L. and Camargo, M. G. (2015) 'GAD: an R package for ANOVA designs from general principles', *Available on CRAN*.

Schmid, S. *et al.* (2017) 'HyRAD-X, a versatile method combining exome capture and RAD sequencing to extract genomic information from ancient DNA', *Methods in Ecology and Evolution*, 8(10), pp. 1374–1388. doi: 10.1111/2041-210X.12785.

Souza, C. A. *et al.* (2017) 'Efficiency of ddRAD target enriched sequencing across spiny rock lobster species (Palinuridae: Jasus)', *Scientific Reports 2017 7:1*, 7(1), pp. 1–14. doi: 10.1038/s41598-017-06582-5.

Sthle, L. and Wold, S. (1989) 'Analysis of variance (ANOVA)', *Chemometrics and Intelligent Laboratory Systems*, 6(4), pp. 259–272. doi: 10.1016/0169-7439(89)80095-4.

Strijk, J. S. *et al.* (2020) 'Museomics for reconstructing historical floristic exchanges: Divergence of stone oaks across Wallacea', *PLOS ONE*. Edited by T. Robillard, 15(5), p. e0232936. doi: 10.1371/journal.pone.0232936.

Suchan, T. *et al.* (2016) 'Hybridization Capture Using RAD Probes (hyRAD), a New Tool for Performing Genomic Analyses on Collection Specimens', *PLOS ONE*, 11(3), p. e0151651. doi: 10.1371/JOURNAL.PONE.0151651.

Timmermans, M. J. T. N. *et al.* (2016) 'Rapid assembly of taxonomically validated mitochondrial genomes from historical insect collections', *Biological Journal of the Linnean Society*, 117(1), pp. 83–95. doi: 10.1111/BIJ.12552.

Untergasser, A. *et al.* (2012) 'Primer3--new capabilities and interfaces', *Nucleic acids research*, 40(15). doi: 10.1093/NAR/GKS596.

De Vere, N. *et al.* (2012) 'DNA Barcoding the Native Flowering Plants and Conifers of Wales', 7(6), pp. 1–12. doi: 10.1371/journal.pone.0037945.

Wandeler, P., Hoeck, P. E. A. and Keller, L. F. (2007) 'Back to the future: museum specimens in population genetics', *Trends in Ecology & Evolution*, 22(12), pp. 634–642. doi: 10.1016/J.TREE.2007.08.017.

White, T. J. *et al.* (1990) 'Amplification and direct sequencing of fungal ribosomal RNA Genes for phylogenetics', (May 2014).

Wood, S. N. (2011) 'Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models', *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1), pp. 3–36. doi: 10.1111/J.1467-9868.2010.00749.X.