Biodiversity Data Journal 11: e102317
doi: 10.3897/BDJ.11.e102317

Methods

Developing the Protocol Infrastructure for DNA Sequencing Natural History Collections

Giada Ferrari[‡], Lore Esselens[§,|], Michelle L Hart[‡], Steven Janssens[¶,#], Catherine Kidner[‡], Maurizio Mascarello[¶], Joshua V Peñalba[¤], Flávia Pezzini[‡], Thomas von Rintelen[¤], Gontran Sonet[|], Carl Vangestel[|], Massimiliano Virgilio[«], Peter M Hollingsworth[‡]

‡ Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom
§ Royal Museum for Central Africa, Tervuren, Belgium
| Royal Belgian Institute of Natural Sciences, Brussels, Belgium
¶ Meise Botanic Garden, Meise, Belgium
# Leuven Plant Institute, Department of Biology, Leuven, Belgium
¤ Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany
« Royal Museum for Central Africa, Department of African Zoology, Tervuren, Belgium

Corresponding author: Peter M Hollingsworth (phollingsworth@rbge.org.uk)

Academic editor: Gergin Blagoev

Received: 18 Feb 2023 | Accepted: 04 Aug 2023 | Published: 27 Oct 2023

Citation: Ferrari G, Esselens L, Hart ML, Janssens S, Kidner C, Mascarello M, Peñalba JV, Pezzini F, von Rintelen T, Sonet G, Vangestel C, Virgilio M, Hollingsworth PM (2023) Developing the Protocol Infrastructure for DNA Sequencing Natural History Collections. Biodiversity Data Journal 11: e102317. https://doi.org/10.3897/BDJ.11.e102317

Abstract

Intentionally preserved biological material in 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. In this paper, we review the practical challenges associated with making the recovery of DNA sequence data from museum collections more routine. We first review key operational principles and issues to address, to guide the decision-making process and dialogue between researchers and curators about when and how to sample museum specimens for genomic analyses. We then outline the range of steps that can be taken to reduce the likelihood of contamination including laboratory set-ups, workflows and working practices. We finish by presenting a series of case studies, each focusing on protocol practicalities for the application of different mainstream methodologies to museum specimens including: (i) shotgun sequencing of insect mitogenomes, (ii) whole genome sequencing of insects, (iii) genome skimming to recover plant plastid genomes from herbarium specimens, (iv) target capture of multi-locus nuclear sequences from herbarium specimens, (v) RAD-sequencing of bird specimens and (vi) shotgun sequencing of ancient bovid bone samples.

Keywords

Museomics, hDNA, biodiversity genomics, natural history collection sequencing

Introduction

Natural history collections as a resource for genomic science

There are more than one billion specimens representing ca. two million species stored in natural history collections worldwide (Wheeler et al. 2012, Yeates et al. 2016, Johnson et al. 2023). These collections span a wide geographical and temporal range and represent a globally distributed biorepository. They house biological specimens representing the world’s known species, along with many specimens representing undescribed species awaiting taxonomic recognition and formal taxonomic descriptions (Bebber et al. 2010). First and foremost, these natural history collections were established to support understanding of species diversity and distributions (Miller et al. 2020) and the vast majority of specimens housed in these repositories were collected to preserve their appearance and morphological features; most specimens were not collected with DNA sequencing in mind (Roycroft et al. 2022).

Until recently, the recovery of DNA sequences from museum specimens was challenging and prone to very high rates of failure or requiring laborious protocols for successful recovery of minimal quantities of nucleotide sequence data (Lalueza-Fox 2022). However, with the development of improved sequencing technologies and protocols, there is now a rapid surge of interest in the field of museomics (Card et al. 2021, Raxworthy and Smith 2021). Considerable attention is being given to unlocking DNA data at a large scale, capitalising on the centuries of effort that have gone into the acquisition of biological specimens for natural history collections (Hebert et al. 2013, Folk et al. 2021). At a very practical level, natural history collections provide access to easy-to-retrieve and often well-identified specimens. This contrasts with the considerable challenges and costs associated with obtaining freshly-collected material for DNA analyses, such as field collecting costs, the cost of preparing voucher specimens and the difficulties of accessing taxonomic expertise to ensure accurate biosample identifications (Hebert et al. 2013, Camacho et al. 2018). These challenges are exacerbated for taxa occurring in remote and/or poorlystudied locations (Wandeler et al. 2007) or areas that are difficult to access because of political instability or conflict (Burrell et al. 2015). Furthermore, where taxa or populations have been lost in the wild, natural history collections are often the only genetic resource for extinct and endangered species (Wandeler et al. 2007, Clewing et al. 2022). Beyond these practical benefits of sampling museum specimens for DNA, there are also the unique scientific opportunities that come from being able to undertake time-series analyses capitalising on the temporal component of natural history collections; DNA sequencing of these collections can provide direct windows into evolutionary processes and patterns of adaptation and evolutionary change (Holmes et al. 2016) and the trajectory of species of conservation concern (Nakahama 2021, Jensen et al. 2022). Likewise, sequencing specimens from natural history collections can also provide insights into the dynamics of associated organisms, such as pathogens, parasites and other intimately-connected species residing in or on museum specimens (Bieker et al. 2020, Ferrari et al. 2020, Ristaino 2020, Raxworthy and Smith 2021, Speer et al. 2022).

Storage and preservation of museum specimens

The global collection of preserved natural history specimens contains a diverse set of samples encompassing a multitude of different tissue types and preservation methods (Carter and Walker 1999). Major collections that are stored dry include pressed plant and fungal herbarium specimens, pinned insects, bones, teeth, shells, skins and hides. Specimens that are stored dry are often subjected to direct heat treatment during the drying process and, in some cases, chemical treatments. For instance, plant material from the Tropics was often immersed in alcohol prior to direct-heat drying to prevent specimens rotting in humid environments (Hodge 1947). Animal skins may have been prepared using a wide range of techniques, including air drying, salting, tanning and chemical treatments, such as arsenic (McDonough et al. 2018). Likewise, a substantial proportion of natural history museum specimens are stored wet, in spirit-based fixatives, including whole specimens, individual organs and other body parts of a diverse array of animals; many of these wet museum collections (especially fish, reptiles and amphibians) were often fixed with, or may still be preserved in, formalin (Hahn et al. 2021).

Properties of DNA in natural history collections and implications for sequencing museum specimens

The DNA within natural history museum specimens has distinct properties from the DNA in freshly-collected material which has practical implications for recovering sequence data. From a biochemical perspective, DNA isolated from natural history collection specimens shares many similarities with ancient DNA (aDNA). Characteristically, aDNA is highly fragmented and biochemically damaged, often present in small quantities and subject to contamination from the environment and human handling. In the absence of the enzymatic repair mechanisms of living cells, DNA is subject to hydrolysis, oxidation and cross-linking (Dabney et al. 2013a), processes that can be accelerated by high temperatures, extreme environmental pH, humidity and the presence of microorganisms (Willerslev and Cooper

2005, Willerslev et al. 2007). Hydrolysis and oxidation can lead to depurination, which results in DNA strand breakage (Lindahl 1993). As a consequence, aDNA is typically no longer than 150 bp (Green et al. 2009).

These degradation processes for aDNA also occur to greater or lesser degrees in museum specimens. Various studies of DNA degradation in natural history collections have shown that DNA fragmentation can occur rapidly after death (Sawyer et al. 2012), with a wide range of reported fragment lengths, including frequent reports of fragment lengths < 100 bp (McDonough et al. 2018, Canales et al. 2022, Mullin et al. 2023). There is an imperfect relationship between specimen age and levels of fragmentation, with some studies showing a correlation between specimen age and fragment length (McCormack et al. 2016 , Weiß et al. 2016, Mullin et al. 2023), whereas others did not (Sawyer et al. 2012). The factors affecting the rate of fragmentation are complex (Kistler et al. 2017) and include differences between genomes (e.g. mtDNA sequences showing slower degradation than nuclear sequences (Heintzman et al. 2014)), differences between tissue types (Kistler et al. 2017, Andreeva et al. 2022) and differences due to the environmental conditions at the site of specimen preparation and different storage environments and preservation methods (Brewer et al. 2019, Card et al. 2021, Mullin et al. 2023). For instance, museum storage conditions may impact on biomolecule degradation, with temperature and humidity influencing levels of preservation (Kistler et al. 2017, Brewer et al. 2019). Likewise different preservation methods themselves impact on the preservation and recoverability of nucleotide sequences: several preparation techniques involve heat, which accelerates DNA hydrolysis resulting in fragmentation (Lindahl 1993, Willerslev and Cooper 2005). Formalin fixation is a commonly-used technique for wet-mounted specimens and, especially if unbuffered, can cause a number of reactions, including DNA fragmentation via acid-driven hydrolysis and DNA-protein cross-linking that results in PCR inhibition (Brutlag et al. 1969, Gilbert et al. 2007b). Treatment of bones with ammonium solutions and various tanning agents has also been reported to reduce DNA yields from museum specimens (Vuissoz et al. 2007, Tarbet Hust and Snow 2021). Finally, pest-control treatments of collection specimens can also impact the recovery of DNA (Espeland et al. 2010, Töpfer et al. 2011).

Following the fragmentation of DNA in museum specimens, there is a consequential and associated loss of DNA as DNA fragments diffuse away from specimens. Kistler et al. (2017) proposed a model by which DNA fragmentation occurs rapidly after death before slowing down, then bulk diffusion leads to the decay of DNA concentration through time. This highlights the importance of tissue types which create closed systems that minimise DNA loss for the retention and recovery of nucleotide sequences (e.g. dense bone tissue, seed tissue).

There are several consequences of this fragmentation and loss of DNA. The first is that experimental effort may be expended which ultimately leads to a failure to recover DNA sequence data due to low endogenous DNA content. The second is that DNA sequence data may be recovered, but be misleading due to contamination. The potential for contamination in sequence data from museum specimens is substantial. The low concentrations of fragmented endogenous DNA in museum specimens represent an initial low signal-to-noise ratio and a high potential for contamination from a wide variety of sources, including:

1. Biological material on the specimen (surface contaminants and biological materials associated with specimen preparation);

2. High concentrations of DNA from fresh samples and their amplification products processed in the same facility which represent an important source of contamination when handling degraded DNA. Such contaminant DNA may be present in higher concentrations than the DNA in historic samples and this is exacerbated by subsequent PCR being biased towards higher-quality DNA;

3. General contamination in the processing lab, including sources of contaminating DNA from specimen handling, laboratory reagents and aerosols in the wider environment.

At best, contamination reduces sequencing efficiency for endogenous DNA and requires greater sequencing efforts at higher costs. What is more problematic is the generation of erroneous data where misleading biological inferences are made from undetected contamination (Yeates et al. 2016).

A more generic source of error, but one to which museum-derived sequences are particularly susceptible, is problems stemming from low coverage of sequence reads due to low DNA concentration. This can result in misleading inference; for example, failure to recover both alleles in diploid heterozygotes leading to an overestimation of homozygosity at some loci and in some specimens (Ewart et al. 2019) or, more generally, the introduction of noise due to miscalls which may simply override any weakly-resolved genuine signal in the data.

Above and beyond the challenges of recovering reliable sequence data from low concentrations of fragmented endogenous DNA is the possibility of post-mortem modifications leading to artefactual substitutions in the recovered DNA sequences. Post-mortem hydrolytic deamination causes base modifications, primarily affecting cytosine (Orlando et al. 2021). Uracil, the deamination product of cytosine, causes the misincorporation of adenine during DNA amplification. This results in C to T substitutions in the deaminated strand and G to A substitutions in the complementary strand of DNA (Briggs et al. 2007, Brotherton et al. 2007). This deamination of cytosine resulting in C to T and G to A substitutions during amplification can, if unchecked, lead to a systematic misleading signal in the data. Sequences from different biological samples may share nucleotide changes due to these miscoding lesions which may be misinterpreted as genuine biological similarities. The accumulation of miscoding lesions at the end of DNA molecules is a feature of aDNA (and used as an important parameter for the validation of aDNA data authenticity (Briggs et al. 2007, Green et al. 2009)); however, they do tend to be less common in studies focusing on natural history collections. DNA deamination correlates with specimen age and there is an expectation for more recently collected museum specimens to show limited impacts of deamination-related substitutions, compared to centuries-old specimens which, in turn, are expected to show substantially lower impacts than those found in ancient DNA material (Weiß et al. 2016, Kistler et al. 2017, Canales et al. 2022).

Outstanding challenges to the routine sequencing of museum specimens

There has been a recent shift in the field of museomics from small-scale studies, often with high rates of failure, to increased success rates and a growth of increasingly-ambitious studies aiming to liberate sequence data at a large scale from museum collections (Hebert et al. 2013, Alsos et al. 2020, Kates et al. 2021, Mullin et al. 2023). The complexity of museum collections themselves and the variation in specimen ages, tissue types, preservation methods and storage conditions, preclude simple universal high-throughput methods. Nevertheless, there is considerable scope for continued optimisation of approaches and community development and dialogue around appropriate and optimal working standards. Within this general challenge, specific areas for development in making the recovery of sequence data from museum specimens more reliable and routine include:

- Deciding when it is appropriate to sample museum specimens: Development of guiding principles to facilitate sampling decisions that support specimen utilisation, but avoid unnecessary and unproductive destructive sampling;

- Minimising the risks of contamination and production of erroneous sequence data: Guidance and utilisation of appropriate laboratory infrastructure and data verification steps;

- Maximising the recovery of endogenous DNA sequences: Optimisation of protocols to improve the efficiency and efficacy of different widely used techniques.

These three topics are addressed in subsequent sections of this paper.

Deciding when it is appropriate to sample museum specimens for DNA sequencing

Destructive sampling poses a dilemma between damaging a specimen for research utilising existing protocols and preserving the specimen for future and improved methodologies (McDonough et al. 2018). Thus, museomic studies need to consider the likelihood of successfully obtaining DNA sequences, the scientific insight that can be obtained, the amount of material needed with regards to specimen and collection size and the potential benefits of postponing sampling to await future methodological advances. Where available, low(er) value specimens in museum collections represent useful material for protocol development and testing prior to destructive sampling on material of higher value. Many museum collections contain samples of limited taxonomic value (e.g. large volumes of sterile material), samples with poor meta-data or specimens which have abundant duplicate material. Such specimens represent more suitable candidates for experimentation than high-value, important and/or unique individual specimens, such as type material.

To minimise damage to specimens, a number of minimally or non-destructive sampling protocols for collection material have been proposed. DNA extraction protocols for ancient and historic DNA have been optimised to obtain good DNA yields from small amounts of material, for example, as low as two milligrams of dry plant tissue (Latorre et al. 2020) or one insect leg (e.g. Cavill et al. (2022)). In terms of less invasive sampling, approaches include sampling the embedding fixative solution of wet collection specimens (Rayo et al. 2022), rubbing an eraser over herbarium specimens (Shepherd 2017) and gentle digestion followed by drying of teeth (Rohland et al. 2004), whole insect specimens (Gilbert et al. 2007a, Korlević et al. 2021) or herbarium material (Sugita et al. 2020). Such approaches can be extremely useful for many applications, although 'non-destructive' sampling often yields lower DNA amounts, limiting genetic analyses to low coverage of DNA sequences or high copy-number regions, such as organellar genes.

Maximising the use of museum specimens, including for genomic analyses, while minimising unnecessary destruction of precious samples thus reflects a balancing act. This is made particularly difficult, as often the greatest scientific returns will come from the specimens that are most valuable. For instance, type specimens will almost always have significant constraints on their use, which may act as a barrier to inclusion in genomic studies. On the other hand, effective minimally destructive sequencing of type specimens provides a direct connection between genomic data and the application of a species name and, hence, represents a significant scientific benefit, particularly for taxonomic and systematic studies. There is a general point, that while sampling a museum specimen for genomic analysis usually results in something being taken away from the specimen to obtain DNA, it can also result in something extremely useful being added, in terms of critically important genomic data which may add considerable value to the specimen (i.e. the concept of the extended specimen; Webster (2017)).

Guidance on best practice standards and processes for the access and transfer of samples for genomic analysis is given by the Consortium of European Taxonomic Facilities (2015) and de Mestier et al. (2022). However, from a perspective of the impacts on the specimens themselves, it is not surprising, given the rapidly-evolving state of the field and the complexity of choices regarding different collections and different approaches, that there are no community standards to guide *when* it is appropriate to destructively sample specimens. There are various policy documents to guide decision-making at institutional levels and several useful more general perspectives (e.g. Freedman et al. (2018), Austin et al. (2019), Pálsdóttir et al. (2019)). To further facilitate the navigation of 'when and how' to sample, we outline ten key principles which can usefully be followed by researchers and assessed by curators in guiding when to undertake destructive sampling of specimens for genomic analyses:

1. Assess the scientific merit of the planned genomic project; ensure there is a clear likely benefit prior to commencing destructive sampling and that the resulting data will be informative and of sufficient resolution to tackle the question at hand;

2. Always adopt a minimally destructive approach for genomic studies of museum specimens unless there is a clear surplus of available tissue, such as extensive duplicate specimens or 'sacrificial' specimens available for experimentation;
3. Utilise alternative options to destructively sampling important specimens if available (e.g. make use of any reliable previously-sampled tissues or previous DNA extracts, adopt a non-destructive sampling approach, if applicable);
4. If multiple tissue types are available, consider the likely success rates for the different tissues and weigh this against their respective morphological impacts on the specimen in choosing which tissue to sample;
5. Seek to maximise the reusability of the data from destructive sampling: consider methodologies which give maximum amounts of data which will be of use for multiple downstream applications;
6. Seek to maximise the reusability of DNA from destructive sampling: adopt methods of handling and storing DNA extracts to maximise their preservation and reuse potential to minimise further need for specimen sampling;
7. Process samples following appropriate laboratory controls and with clear data-verification steps to ensure that the resulting data have maximum reliability and value;
8. Evaluate the feasibility of success prior to destructive sampling of valuable specimens in terms of protocol efficacy, researcher capability and laboratory suitability and only proceed where the likely chances of success and resulting scientific benefits outweigh the costs of any destructive sampling;
9. Report successes and failures to guide future optimisation of protocols and decision-making regarding destructive sampling;
10. Ensure appropriate accessibility of the resulting sequence data and linkages and connections between the data and the specimens they were derived from, to ensure that specimen sampling for genomic analyses results in added value to the specimen itself.

Minimising the risks of contamination and generation of erroneous sequence data

Historic DNA versus ancient DNA versus modern optimally-preserved DNA

The opportunities arising from the sequencing of museum specimens have attracted researchers from different backgrounds and fields. On the one hand, sequencing the degraded DNA in museum specimens has long been a focus for aDNA researchers. Ancient DNA techniques involve working with low concentrations of highly-degraded DNA in specialist laboratories with strict guidelines and meticulous anti-contamination precautions (Gilbert et al. 2005, Llamas et al. 2017). These techniques and working practices developed for aDNA allow the recovery of genetic material up to and over one million years old (van der Valk et al. 2021, Kjær et al. 2022). On the other hand, taxonomists, systematists and other researchers focusing on contemporary biological samples often process larger numbers of samples in more general laboratory facilities working with tissue samples preserved in a fashion aimed at maintaining high concentrations of non-degraded DNA.

The DNA in the majority of natural history museum specimens sits at the interface of aDNA and non-degraded DNA samples and is classed as historic DNA (hDNA), more formally defined as DNA from specimens archived in museum collections that were not originally intended as genetic resources (Billerman and Walsh 2019, Raxworthy and Smith 2021, Irestedt et al. 2022). This classification recognises that museum specimens, typically collected over the last 250 years, have different properties to specimens recently collected and preserved for DNA analyses and ancient specimens deposited in nature over millennia (Raxworthy and Smith 2021, Wandeler et al. 2007). This distinction between ancient and historic DNA is useful, although it should be noted that the line between archaeological specimens, natural history collections and even biobank material is a blurred one. As noted previously, several factors other than age influence DNA preservation, such as temperature, substrate, taphonomic conditions and specimen preparation and storage. Thus, a permafrost-preserved archaeological sample may be a better DNA source than a heat-dried or chemically-treated museum voucher. Because of the lack of *a priori* information regarding the magnitude of DNA degradation in historical collection material, a pragmatic working assumption for hDNA material is to assume damage and fragmentation, as well as environmental contamination (Latorre et al. 2020).

Current operational practices for processing hDNA

The maximally effective recovery of hDNA is dependent on determining the appropriate levels of stringency of laboratory practices, which minimise risks of contamination, while at the same time, being sufficiently scalable to allow maximum utilisation of the vast resources of specimens available in museum collections. Thus, while utilising dedicated aDNA facilities (Fulton 2012) and the full suite of operational precautions for processing aDNA material is the conservative option, it will not be appropriate in all cases (Raxworthy and Smith 2021) and full aDNA facilities and protocols for all specimens would represent a substantial constraint on sample processing.

During the preparation of this paper, discussions amongst the authors and an informal survey of colleagues working in a range of organisations involved in sequencing museum specimens, revealed a wide range of operational practices. These ranged from processing samples in the same laboratories as fresh tissue, through to dedicated hDNA (or low-copy) laboratories, through to only ever using fully-equipped aDNA facilities for processing museum specimens. A multitude of factors were articulated as underlying the decision-making of which facilities to use for processing hDNA samples, including:

- Resource constraints (money and/or space) precluding establishment of a dedicated facility;
- Desire to use existing facilities in standard labs to enable processing of large numbers of samples;

- Controls in place for data authentication and/or stringent cleanliness conditions in standard labs considered adequate to negate the need for a dedicated facility for hDNA samples;
- Individual preferences of researchers determining where samples are processed without clear institutional policies;
- When both aDNA and hDNA samples have to be processed in the same institution, the aDNA laboratory being used exclusively for aDNA samples, with museum specimens processed elsewhere due to concerns that the higher concentrations of DNA from museum specimens may lead to contamination problems in the aDNA lab.

Laboratory set-ups and workflows for hDNA sequencing

A general observation noted by several researchers familiar with working with non-degraded DNA samples was a lack of clarity over what an optimal laboratory set-up would look like for hDNA analysis. To facilitate evaluation of options for contamination control and the practicalities of laboratory set-ups, we outline below the headline infrastructure and working practices of an aDNA facility and a basic hDNA facility. We also list contamination-limiting recommendations for processing degraded material in existing, more general, laboratories.

Ancient DNA facility

The most critical component of setting up an aDNA laboratory (Gilbert et al. 2005, Fulton 2012, Knapp et al. 2012, Llamas et al. 2017) is a strict separation of pre- and post-PCR areas. This includes no movement of equipment, reagents, consumables or samples from post- to pre-PCR. Similarly, scientists should not move from the post- to pre-PCR without showering and changing clothes. The dedicated aDNA pre-PCR facilities are physically isolated from any post-PCR area and are used for sample processing, DNA isolation and setting up of sequencing library and PCR reactions. These reactions are moved to post-PCR facilities at the first DNA amplification step and post-amplification products can be handled normally in shared laboratory facilities. Amplification products, fresh biological material and modern DNA samples should never be introduced into the aDNA facilities. Additionally, the dedicated aDNA laboratory should be fitted with a positive pressure, HEPA-filtered air system and UV lights for daily sterilisation of all surfaces. Equipment, tools and working surfaces should be cleaned daily or after every use with a 1-2% sodium hypochlorite solution or a surface decontaminant such as DNA Away™ or DNA Exitus™. Plastic consumables should be UV-sterilised and only filter tips should be used. Everything that is introduced into the laboratory should be decontaminated. The aDNA laboratory should be accessed through an antechamber where incoming reagents and consumables are sterilised (and ideally introduced through a dedicated UV-hatch) and PPE donned (overalls with hood, hairnets, facemasks, face shields, double gloves and shoe covers or dedicated shoes). Destructive sampling and sample powdering should take place in a separate room inside a PCR cabinet or dead-air box. Samples and DNA extracts should be handled exclusively in HEPA-filtered laminar flow cabinets, equipped with UV lamps (and should be cleaned and sterilised after every use). A separate cabinet for DNA-free applications (aliquoting reagents, preparing master mixes) is needed. Additional good working practices include processing samples in small batches and the inclusion of several non-template negative controls for DNA isolation and library preparation, as well as dividing reagents into smaller aliquots.

Historic or low-copy DNA facility

Natural history collection material can be handled in aDNA laboratories. However, aDNA laboratories and their upkeep can be prohibitive in cost; therefore, institutions working exclusively on natural history collections may choose a less stringent set-up for a dedicated hDNA pre-PCR facility. This would usually be located in existing rooms rather than a purpose-built laboratory. Not requiring a positive pressure air system and a laboratory antechamber allows for more flexibility in the choice of location (e.g. repurposing of laboratory spaces) and significantly reduces the costs. It should be noted that if an hDNA facility is being established by repurposing an existing laboratory, thorough cleaning with sodium hypochlorite of all surfaces is essential and new, dedicated equipment should be bought. Similarly to aDNA facilities, all work should take place inside UV-fitted PCR cabinets, and destructive sampling and sample powdering should take place in a separate room or at least in a separate cabinet. Additional UV lamps for surface decontamination may be fitted; however, repeated UV exposure is damaging to laboratory equipment, increasing upkeep costs. Cleaning routines and good practices as described for aDNA facilities should be implemented or adapted as best as possible, most importantly the separation of pre- and post-PCR working areas. In Suppl. material 1, we outline a relatively inexpensive (ca. €86K) and pragmatic equipment list for establishing a dedicated low-copy facility for hDNA processing, that is flexible enough to be installed without extensive building works.

Contamination-limiting measures for working in existing facilities

Due to space or financial constraints or because of the need for higher throughput, it remains the case that many institutions may decide against an aDNA or hDNA facility and process natural history collection material in existing laboratories alongside fresh biological material. Fresh material and especially its amplification products, represents a common source of contamination for historic material. Thus, the separation of pre- and post-PCR areas remains essential, although the routes to achieving this can be varied. At the very least, thermocyclers should always be located in the post-PCR area and movement of samples, reagents, consumables and equipment between pre- and post-PCR should be avoided or limited as best as possible. Additionally, pre-PCR work on collection material should be carried out in dedicated laminar flow hoods (to be cleaned and UV-sterilised regularly) with dedicated tools and reagents.

Data verification and contamination controls

In addition to the physical layout of laboratories, data verification and contamination controls include:

Verification checks against reference libraries: The likelihood of being misled by contamination in the sequencing of museum specimens is a function of the stringency of the controls and the complexity of the detection task. Data-authentication steps can be relatively straightforward, where there is an *a priori* expectation of the sequence to be recovered and an existing reference resource to check it against. Thus, studies like large-scale DNA barcoding projects, are intrinsically well-suited to contamination checks, with small regions of DNA being recovered from individual museum specimens which are usually identified to species level and whose identity can be checked by sequence cluster placement in existing DNA barcode reference libraries such as BOLD (Ratnasingham and Hebert 2007, Ratnasingham and Hebert 2013) if there is sufficient coverage of the study taxon. Likewise, genome skimming studies or target capture-based recovery of organelle genomes can benefit from systematic comparisons of extracted barcode loci against barcode reference libraries (Timmermans et al. 2016, Alsos et al. 2020), as well as wider checks against the growing reference datasets of organelle genomes (e.g. Li et al. (2021)). The data verification checks become more challenging where the density of reliable reference data for comparison is lower and the complexity of the sequence data produced is higher. There are various major infrastructure projects underway driving the development of reference libraries which will facilitate these checks. The International Barcode of Life Project (iBOL) continues to accelerate the production of standardised DNA barcode data. BIOSCAN, the current phase of the iBOL project, aims to generate barcode coverage for two million species by 2027 (Hobern 2021) and the associated development of national and regional barcode initiatives (such as BIOSCAN Europe https://www.bioscaneurope.org/) have a strong focus on the production of tightly-curated reference barcode libraries. Another complementary large-scale biodiversity genomics infrastructure project, the Earth Biogenome Project, has the goal of producing reference genomes for all eukaryotic species (Lewin et al. 2022) and various large-scale geographically-focused projects are underway (The Darwin Tree of Life Project Consortium 2022); the increased density of high-quality reference genomes will greatly support data verification and assembly of the short-read sequences from museum specimens as is already possible with human DNA (de Filippo et al. 2018).

Ordering samples to avoid closely-related taxa being in adjacent wells: The most difficult contamination to spot is from closely-related samples, as even detailed analysis and comparisons with reference samples may not flag contaminants. Where there is a mixture of closely- and more distantly-related taxa being processed, a simple option is to order samples to maximise the likelihood of adjacent well contamination being detectable, by minimising the presence of closely-related specimens in adjacent wells.

Negative controls: The inclusion of non-template negative controls at the DNA isolation and library preparation/PCR step is essential for ensuring that the pre-PCR facilities and reagents are sufficiently clean. *Negative controls* should also be taken through all post-amplification steps, sequenced and included in the data analysis.

Sequencing strategies: Jumping PCR (Kircher et al. 2012) and index switching during cluster generation on the Illumina sequencing platform (van der Valk et al. 2020) can result in reads potentially being assigned to the wrong sample. This normally does not pose a problem for libraries generated from high-quality DNA, but can introduce an artefactual contamination into degraded DNA libraries. This can be overcome by unique dual-indexing of libraries, a common practice in aDNA experiments. Using unique indices in both library adapters allows the detection and removal of these chimeric PCR products. Unique dual-indexing, if not repeated within the same laboratory, also allows monitoring of potential cross-contamination between projects.

DNA repair: To mitigate against misleading inference due to post-mortem DNA modifications, enzymatic treatment of DNA extracts can be undertaken, for instance with the USER reagent (New England Biolabs), a mix of uracil–DNA–glycosylase (UDG) and endonuclease VIII (Orlando et al. 2021). On the other hand, if post-mortem DNA modifications are of importance for authentication (e.g. for aDNA studies), then avoidance of this step is equally important.

Data authentication and validation: In addition to checks against reference libraries, various bioinformatic pipelines and workflows support the verification and authentication of degraded DNA sequences. These include estimating contamination by screening for unexpected results such as 'heterozygosity' of haploid genomes (mitogenomes, plastomes) (Krause et al. 2010, Renaud et al. 2019). These approaches, based on deviations of expected ploidy, require sufficiently high sequencing coverage, but do not necessitate any *a priori* knowledge of the origin of the contamination (Peyrégne and Prüfer 2020). Likewise, testing for the presence of sequencing artefacts due to post-mortem DNA damage (C to T and G to A misincorporations) at the read ends can be undertaken using specific software such as mapDamage2 (Jónsson et al. 2013) and PMDtools (Skoglund et al. 2014). This is particularly relevant to older samples.

Maximising the recovery of endogenous DNA sequences

Over the last decade, there has been a constantly expanding set of literature outlining new developments which contribute to making the recovery of sequence data from museum specimens more cost-effective and routine (Knyshov et al. 2019). These include breakthroughs in sample and tissue types that were previously considered intractable, including significant improvements in prospects for recovery of genomic data from formalin-fixed tissues (Ruiz-Gartzia et al. 2022). For instance, Straube et al. (2021) used aDNA extraction protocols and single-stranded DNA library preparation to achieve high rates of success in recovering endogenous DNA from wet museum collections of a range of vertebrate taxa including formalin-fixed samples, followed by a target-capture approach to recover almost complete mitogenomes from a subset of these samples. Likewise, Hahn et al. (2021) used a hot-alkaline lysis approach for DNA extraction, followed by whole genome sequencing to successfully recover mitochondrial genomes and up to 3X nuclear genome coverage from a diverse range of formalin-preserved vertebrate tissue specimens and outlined a framework to guide decision-making for genomic studies utilising spirit-preserved collections.

The generation of guidelines and decision-making frameworks to support more routine recovery of sequence data from collections is of considerable value as the field of museum collection sequencing expands (McDonough et al. 2018). Of particular use are very large-scale studies processing thousands of specimens which offer the potential for general predictions to emerge regarding the likelihood of recovery of useful sequence data. For instance, Kates et al. (2021) processed nearly 8000 herbarium specimens from a range of angiosperm families from six different herbaria in the United States and evaluated factors influencing DNA recovery and sequencing success using a target-capture approach. They showed the strongest predictor of success was related to taxonomic group (some families performed better than others, likely due to physical or chemical properties of the samples). There was a more limited correlation between DNA yield or sequencing success on the one hand, and the age of specimens or their greenness on the other ('greenness' is an indicator of the preservation process as green specimens are less likely to have been subject to heat or ethanol treatment during drying). Likewise, extensive studies processing thousands of preserved insect specimens from museum collections for DNA barcoding have developed efficient workflows that recover cytochrome oxidase 1 (CO1) sequences for many specimens using primer cocktail sets and Sanger sequencing for younger specimens, coupled with more intensive multiplex PCR and high-throughput sequencing platforms to recover barcode sequences from older specimens (Prosser et al. 2016, Levesque-Beaudin et al. 2023, Santos et al. 2023).

Table 1 highlights a selection of studies outlining recent progress, breakthroughs and protocol developments which support the more routine recovery of genomic data from museum specimens.

Table 1.

Selected papers outlining recent progress, breakthroughs and protocol developments that support the more routine recovery of genomic data from museum specimens

Study taxon	**Tissue type**	**Specimen ages (yrs)**	**Approach(es)**	**Key finding**	**Reference**
Plants	Dried herbarium specimens	2-182	Multi-locus nuclear sequence capture	Large-scale study of 7608 specimens using angiosperm 353 target capture baits; DNA yield poor predictor of sequencing success, plant family strongest predictor of success, successful recovery of old specimens from tropical climates	Kates et al. (2021)

Study taxon	Tissue type	Specimen ages (yrs)	Approach(es)	Key finding	Reference
Plants	Dried herbarium specimens	NA	NA	Protocols and best practices for working with ancient and historical plant DNA. Includes laboratory set-up, DNA isolation, sequencing library preparation and bioinformatic analyses	Latorre et al. (2020)
Plants	Dried herbarium specimens	Up to 280	Shotgun sequencing	DNA in herbarium specimens degrades faster than in ancient bone. Both fragmentation and deamination accumulate over time	Weiß et al. (2016)
Plants	Dried herbarium specimens	most < 20 yrs, 165 samples > 50 yrs; oldest 153 yrs	Shotgun sequencing	Large-scale genome skimming study of 2051 herbarium specimens recovering plastome and rDNA sequences including standard plant barcodes	Alsos et al. (2020)
Fungi	Rust fungi on dried herbarium specimens	Up to 187	Amplicon-based rDNA sequencing	Protocol development and application to track dynamics of plant pathogens through time sampled from herbarium specimens	Bradshaw et al. (2023)
Fungi	Fungarium specimens	Less than 20	Whole genome sequencing	Generation of draft genome assemblies possible and of value for enhancing resolution of fungal phylogeny	Dentinger et al. (2016)
Birds	Avian skins	up to ca. 150	Whole genome sequencing	Step-by-step guide to workflow and protocols, including steps taken to minimise risks of contamination	Irestedt et al. (2022)
Mammals (mephitids, rodents, marsupials)	Dried museum skins	50-120	Shotgun sequencing	Comparison of DNA yields and bacterial contamination levels in commonly-sampled museum mammalian tissues (bone, claw, skin and soft tissue) and implications for sampling *strategies*	McDonough et al. (2018)
Mammals (grey wolf)	Dried museum skins	90 - 146	Shotgun sequencing	“Single-tube” DNA library preparation methods including adaptations for degraded DNA increase library complexity, yield more reads that map uniquely to the reference genome and reduce processing time compared to other Illumina library preparation methods	Carøe et al. (2018)
Mammals (bison and horse)	Bone	Up to 40,680	Shotgun sequencing	A more accessible single-stranded genomic library preparation method optimised for aDNA	Kapp et al. (2021)

Study taxon	Tissue type	Specimen ages (yrs)	Approach(es)	Key finding	Reference
Mammals (dogs and mammoths)	Bone	Up to 37,080	Shotgun sequencing	Competitive mapping of raw sequencing data to a concatenated reference composed of the target species' genome and the genome of possible contaminants contributes to filtering out contamination from ancient faunal DNA datasets with limited losses of true ancient data	Feuerborn et al. (2020)
Mammals (Cricetidae, rodents, deer mouse)	Frozen liver tissue	17 to 41	Whole genome sequencing	Linked-read or “synthetic long-read” sequencing technologies provide a cost-effective alternative solution to assemble higher quality de novo genomes from degraded tissue samples	Colella et al. (2020)
Insects (Phylinae; plant bugs)	Abdomen	1 to 54	DNA bait capture	Inexpensive data generation to produce sufficient amount of data to assemble the nuclear ribosomal rRNA genes and mitochondrial genomes	Knyshov et al. (2019)
Insects (Apidae, bumble bees)	Leg	18 to 131	Shotgun sequencing	DNA in entomological specimens in NHC highly degraded, process age dependent with a roughly linear reduction in fragment length over time after strong initial fragmentation	Mullin et al. (2023)
Insects (Culicidae, mosquitoes)	Whole specimens	33 to 84	Shotgun sequencing	Minimally damaging extraction method for building libraries for Illumina shotgun sequencing	Korlević et al. (2021)
Insects (Lepidoptera)	Leg	Average 50 years	Sequel CO1 barcoding	Demonstration of utility of PacBio Sequel platform for recovery of full length CO1 barcodes	D'Ercole et al. (2021)

A general challenge for the effective recovery of endogenous DNA from museum specimens is the frequent low complexity of libraries caused by PCR and cleaning steps modifying the relative abundances of the original DNA fragments during library preparation (Casbon et al. 2011). This leads to the formation of artefactual PCR duplicates that may bias sequencing results, decrease final coverage and increase sequencing costs (Rochette et al. 2023). PCR duplicates can be removed during bioinformatic analysis (Marx 2017); however, given the low concentration and quality of DNA from museum specimens, it may be valuable to enhance the library complexity prior to sequencing. To this end, amounts of starting material can be increased where possible (Fu et al. 2018) and single-tube library preparation methods can be used (e.g. Carøe et al. (2018), Kapp et al. (2021)). PCR conditions can also be adapted, for example, by selecting polymerases that are suited to copy degraded DNA templates with good fidelity and without a severe tendency to preferentially amplify DNA templates that are shorter or with higher GC content (Aird et al. 2011, Dabney and Meyer 2012, Seguin-Orlando et al. 2015). Protocols for archival specimens generally perform amplifications in several independent PCR reactions with minimal numbers of cycles (van der Valk et al. 2021, Irestedt et al. 2022) and with sequencing efforts proportional to library complexity (Daley and Smith 2014). Finally, sequencing by synthesis with 50 to 150 cycles format and single-end mode is usually more cost-effective with short degraded fragments (Raxworthy and Smith 2021); however, choosing more cycles or a paired-end mode may be valuable if fragment length distributions are used to filter out contaminants.

Case study overview

To provide further details on protocol development for particular genomic approaches and their success in application to different taxonomic groups, we present a series of case studies undertaken at different institutions as part of the EU SYNTHESYS+ project. Each focuses on protocol practicalities for the application of different mainstream methodologies to museum specimens including: (i) shotgun sequencing of insect mitogenomes (Museum für Naturkunde Berlin), (ii) whole genome sequencing of insects (Royal Museum for Central Africa), (iiii) genome skimming to recover plant plastid genomes from herbarium specimens (Meise Botanic Garden), (iv) target capture of multi-locus nuclear sequences from plants (Royal Botanic Garden Edinburgh), (v) RAD-sequencing of bird specimens (Royal Belgium Institute for Natural Sciences) and (vi) shotgun sequencing of ancient bovid bone samples (Royal Belgium Institute for Natural Sciences).

Case study 1: Assessing the potential of rapid shotgun sequencing for the recovery of mtDNA genomes from pinned insect specimens

Introduction

The world's entomological collections hold more than half a billion pinned (dried) insect specimens (Short et al. 2018), representing an enormous genetic and genomic resource, for tackling a wide range of questions of great scientific and societal importance, not least of which is better understanding global insect declines (Card et al. 2021). For most species-rich insect taxa with a fundamental role in terrestrial ecosystems (pollinators, food sources, parasitoids, primary consumers etc.), reliable diversity estimates for megadiverse tropical regions are at best available for just a few subgroups. By mobilising sequence information (especially from types) from natural history collections, it will become much easier to identify new species as well as synonyms, and thus speed up the process of biodiversity discovery, as well as building a database for DNA-based species biomonitoring.

This potential has received increasing attention in recent years (Raxworthy and Smith 2021) and several studies have looked at the recovery of DNA with a specific focus on DNA isolation protocol optimisation to increase recovery of sequence data and minimise damage to specimens for NGS sequencing (e.g. Patzold et al. (2020), Korlević et al. (2021)). In many instances, even a limited amount of mtDNA data, such as short DNA barcodes, will be sufficient for resolving taxonomic issues and characterising patterns of species diversity in highly-diverse insect taxa (Yeo et al. 2020). Here, we evaluate the performance of shotgun sequencing (low-coverage genome sequencing) of museum specimens with a wide age range from three key insect taxa to explore the utility of a fast, generic and inexpensive approach to obtain mtDNA data from natural history collections to support DNA-based biodiversity inventories and biomonitoring.

Materials and Methods

Specimens (n = 70 in total) were selected from a genus (or two closely-related genera) from each of three major holometabolan insect groups (Diptera, Hymenoptera, Lepidoptera) with a collecting date ranging from 1891 to 2015 (132-8 years in collection storage). The taxa (Diptera: *Sarcophaga*; Lepidoptera: *Eudonia*, *Scoparia*; Hymenoptera: *Xylocopa*) are actively studied by the respective curators at Museum für Naturkunde and the availability of at least one reference mitogenome was an additional core criterion. One leg each was used for DNA isolation using the Qiagen Investigator Kit. Multiplexed libraries for paired-end Illumina sequencing were prepared for each sample with the “NEXTflex Rapid DNA Seq Kit 2.0” and “NEXTflex HT Barcodes” (Bioscientific/PerkinElmer). DNA input varied between 1 and 300 ng (10 ng or less for > 75% of samples). As the DNA of all samples was expected to be strongly degraded, the protocol was adapted accordingly (no shearing and bead-size selection, adjustment of bead buffer - sample ratio for clean-up after adapter ligation). Library quality, size and quantity were determined with a TapeStation (D1000 Kit Kit/ Agilent) and Qubit 2.0 Fluorometer (dsDNA HS Assay Kit/ Thermo Fisher Scientific). Libraries were pooled, based on these parameters and low-coverage test sequenced on an Illumina Miseq (PE150). Based on the results from the test sequencing, 12 libraries were subjected to a booster PCR to increase quantity. Libraries were re-pooled, based on library parameters (concentration) and test sequencing results (number of reads) and the resulting pool was sequenced on an Illumina Nextseq (PE150). The cleaned and de-multiplexed reads were mapped against mtDNA genomes of the respective target taxon using the MITObim pipeline (Hahn et al. 2013).

Results and Discussion

The amount of recovered DNA ranged from 3 ng to 1.2 µg (Fig. 1a). The highest amount of DNA for samples collected before 1950 was 75.6 ng and all samples yielding more than 100 ng were collected after 1950, but note that 50% of these younger samples did not exceed 100 ng either. The overall highest amounts of DNA were recovered from samples collected between 1990 and 2000. This pattern was also recovered when comparing read number and sample age (Fig. 1b) and apart from one exception, all samples with > 2 mio. reads were collected after 1970 and > 10 mio. reads were only exceeded in the two youngest samples (2015). Mapping reads against mitogenome sequences of the target taxa obtained from GenBank recovered a minimum of 14,652 bp for *Xylocopa* (Hymenoptera), 15,299 bp for *Eudonia* (Lepidoptera) and 15,667 bp for *Sarcophaga* (Diptera). The published mitogenome of *Xylocopa* is partial with only 14,655 bp, which might explain the shorter assemblies for that genus. Completeness and coverage varied strongly amongst samples. All samples with a completeness below 50% (7-33%, n = 8; Fig. 1c) were collected before 1950, while no sample collected after 1950 had less than 70% completeness. The pattern for coverage was different, as both a relatively low coverage of < 50 and a high coverage > 100 were found throughout the entire age range of samples (Fig. 1d). An extremely high coverage of > 1000 was only found in *Sarcophaga*.

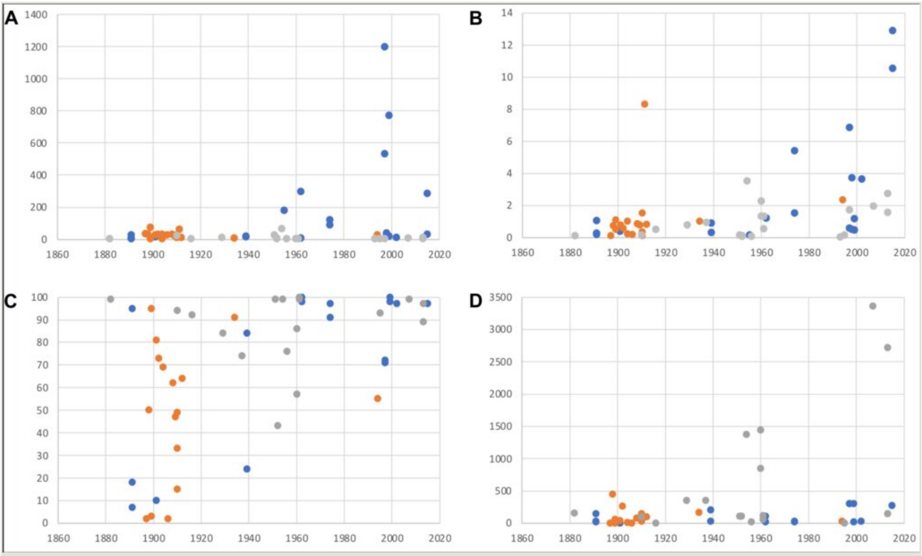

Figure 1. doi

Scatter plots of DNA and sequence recovery from pinned insect specimens by age and taxon (blue - *Eudonia*/*Scoparia*, grey - *Sarcophaga*, orange - *Xylocopa*). Specimen age (in years) is on the x axis in all panels. **A** Total DNA yield (ng). **B** Number of sequencing reads. **C** Completeness of the mitogenomes (%). **D** Coverage (n).

Success rate in relation to sample age

DNA degradation is affected by several variables, such as initial preservation of samples and storage conditions, which in themselves are highly diverse in many aspects (temperature, humidity, pest control chemicals). As the effect of these factors accumulates over time, an obvious assumption is that DNA degradation will be worse in older specimens and sequence recovery consequently more difficult, as has been shown in a recent systematic study using NGS methods (Mullin et al. 2023). Our results support this generalised conclusion to some degree: the oldest samples (collected before 1950) yielded consistently low amounts of DNA and only one of the old specimens yielded more than 2 mio. sequence reads. However, the mixed patterns for (relatively) younger samples support the notion that the initial preservation of samples is also contributing to the degradation of DNA, as about 50% of specimens collected after 1950 yielded about the same magnitude of DNA and sequencing reads as the older samples, which is notable given that some of the poorly-performing younger samples were collected as recently as the year 2000. A rapid degradation of DNA following death was also found by Sawyer et al.

(2012) and Kistler et al. (2017). The completeness and coverage of recovered mitogenomes showed a relationship with sample age, albeit with high levels of variation. Thus, although all samples with low completeness were older than 80 years, there was substantial variation in the completeness of mitogenomes amongst the oldest samples (collected < 1920).

Overall, no sample failed entirely as measured by obtaining reads from mtDNA. One sample did not yield any reads at all after the booster PCR, but the same sample worked (albeit with only ca. 81,000 reads) for the library prepared without additional PCR amplification.

In contrast to age, variation in success rate does not seem to differ amongst taxa, if variation amongst collecting dates is taken into account. Almost all samples of *Xylocopa*, for example, were collected before 1940 and about 75% in 1912 or before, and the apparently lower success rate for *Xylocopa* could be explained by age alone. Similarly, the species and consequently specimens used in this study differed in body size, which was not controlled for here and, as the total amount of DNA recovered is expected to be dependent on tissue input, that value cannot be directly compared meaningfully between specimens of different taxonomic groups.

Effectiveness in terms of costs and time of shotgun sequencing

Shotgun sequencing is a straightforward and technically undemanding approach by NGS standards. In recent years, it has also become ever more cost effective as measured by cost per bp. For example, in this study, each sample yielded on average 1.6 mio. reads and sequencing costs per sample were ca. €32 (price as of December 2022). These costs could be further reduced using different sequencing platforms (e.g. Illumina Novaseq).

Interestingly, to date, a limited number of studies have used shotgun sequencing for museomics in insects and only one has targeted mtDNA in particular for taxonomy in a specimen-based approach. All types of Australian prionine longhorn beetles were shotgun sequenced by Jin et al. (2020), leading to a major revision of their taxonomy. The methodological study by Timmermans et al. (2016) aimed to show the applicability of shotgun sequencing in museum specimens, but by a metagenomic approach that combined DNA extracts without individual sample indexing. Cong et al. (2021) used a shotgun approach for taxonomic purposes in North American butterflies on a large number of specimens, but targeted mainly nuclear genes and their study relied on creating a reference genome from a modern sample first. Other shotgun sequencing studies have aimed to place enigmatic taxa on the tree of life (Twort et al. 2021) or explore population genetic structure (Cridland et al. 2018) or species conservation (Mikheyev et al. 2017) by focussing on deeper sequencing of few (2-29) museum specimens. While these studies focused on aspects of species biology, Mullin et al. (2023) focussed on > 100 specimens of one species of bumble bee from the UK to study DNA preservation in museum specimens. Their results also show that there is great potential in the use of museum specimens of pinned insects for biodiversity genomics research.

A frequently-used alternative to shotgun sequencing for recovering genomic data from museum specimens is target capture (e.g. Blaimer et al. (2016), Mayer et al. (2021)). It has mostly been used for generating substantial amounts of nuclear data, but as yet not so much for mitogenomics (but see Knyshov et al. (2019)). An advantage of target capture is the significant increase in sequencing efficiency, as target genes will make up a much larger proportion of reads, allowing the pooling of more samples and consequently bringing down sequencing costs per sample. The downside of this approach is that it only works reliably up to a genetic distance of ca. 12% between target DNA and bait sequences. Custom bait sets need to be created at a rather low taxonomic level (generally genus or even species groups, depending on genetic divergence between species). This is a more acute problem for mtDNA capture, as genetic divergence in mtDNA is about four times higher than in nuclear DNA. Baits can either be ordered custom made, which is expensive (ca. €120/reaction as per manufacturer's instructions or between €5-15 if 8-24 libraries are captured with the amount of baits delivered for one reaction) or PCR generated (Knyshov et al. 2019). While the latter approach is more cost-effective, with costs of €25-39 in total (including sequencing per sample according to Knyshov et al. (2019)), it does add to the workload.

For both low-coverage shotgun sequencing and target capture, there is a trade-off between costs, time and sequencing success (measured by the completeness of the target sequence(s)), which is likely to tilt towards bait capture when the aim is to sequence a large number of closely-related samples. However, if the aim is to target a larger range of taxa at the genus level and beyond for taxonomic purposes, such as DNA barcoding, then shotgun sequencing has the edge in our opinion due to its relative ease and the universality of approach. As sequencing costs are still on a downward trajectory, the cost balance is likely to be tilted further in its favour in the future.

Main findings and recommendations:

- A shotgun approach is particularly appropriate for obtaining (mtDNA) data for a wide range of different taxa with relatively little effort in the lab, which makes it highly useful for taxonomy and providing reference sequences from type material. If the aim is to generate complete datasets from many individuals of closely-related species, bait capture might be a viable alternative;
- Older samples will often require more sequencing effort to obtain the same amount of data as more recent specimens. If the main aim is the generation of DNA barcodes for taxonomic purposes, this should not be overly relevant in practice;
- When using a shotgun approach, using a leg is sufficient to obtain an adequate amount of data for taxonomic purposes at least from medium-sized to large (> 10 mm) specimens, which should make it easier for curators to give permission for destructive sampling;

- Crucially, when adding to collections, sample preservation should be optimised in the field in order to avoid heavy DNA degradation before specimens become museum specimens.

Data availability: The raw sequencing data from this case study has been deposited at the European Nucleotide Archive (ENA) under project PRJEB59182 with accession IDs ERS14475133 - ERS14475206.

Case study 2: Genomic vouchering in insect museum collections: the quest for a pragmatic approach to routine, large-scale genotyping

Introduction

The costs directly related to genomic library preparation and sequencing represent one of the main limiting factors hampering the whole genome sequencing (WGS) 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 of the past few years have 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 (e.g. Crampton-Platt et al. (2016), Malakasi et al. (2019), Strijk et al. (2020)). In this case study, we present a feasibility study on the routine collection of genomic data from insect museum vouchers. Here, with "routine genotyping" we refer to standard and commonly-used wet-lab pipelines for WGS rather than to the more elaborate, expensive and technically-challenging protocols which are used to recover highly-degraded DNA (reviewed in Orlando et al. (2021)).

Materials and Methods

Comparative performances of commercially-available DNA extraction kits

The performance of commercial DNA extraction kits (see below) was compared in a pilot study targeting the Royal Museum of Central Africa (RMCA) collections of “true” fruit flies (Tephritidae, Diptera) and African hoverflies (Syrphidae, Diptera). We selected three to six specimens from seven sample groups dating from 2008 to 2016. These included three Tephritidae (*Zeugodacus cucurbitae* Coquillett, *Bactrocera dorsalis* Hendel and *Dacus bivittatus* Bigot) and two Syrphidae (*Eumerus* sp. and *Ischiodon aegyptius* Wiedemann) species; all specimens were stored in 100% ethanol at -20°C, except *Ischiodon aegyptius* which was pinned and preserved at room temperature. Digestions in lysis buffers were implemented on whole bodies for all specimens. For comparative purposes, we also processed forelegs of a separate set of specimens. The lysates obtained from each specimen were divided into four aliquots and the DNA purified using spin columns from the

DNA extraction kits listed in Table 2 following the manufacturers' instructions. The experimental design was based on 30 whole specimens and 18 legs (two negative controls were also included); these samples were 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 2.

An overview of the DNA extraction kits tested on fruit flies and hoverflies.

QIAGEN kit (50 samples)	**Spin column**	**Range of DNA fragment sizes (according to manufacturer's instructions)**	**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

Relationships between voucher DNA quality and WGS performance

To assess the relationship between WGS performance and (a) voucher age and preservation and (b) DNA quality and quantity, we targeted a total of 732 insect vouchers archived in the collections of RMCA collected from 1997 to 2022 (Fig. 2) and preserved either in ethanol at -20°C (n = 651), pinned at room temperature (n = 14) or dried DNA stored at room temperature (n = 67). All DNA extractions were performed using the DNeasy Blood and Tissue Kit (Qiagen catalogue nr. 0148945380). We quantified the amount of DNA extracted as measured by a Qubit 4 fluorometer (HS DNA Kit, Thermo Fisher Scientific) and the quality of DNA via DNA fragment size distributions as measured using the DNF-930 dsDNA Reagent Kit (75 bp – 20000 bp) on the fragment analyser of the Genomics Core (Leuven, Belgium). These vouchers were considered as suboptimal for genomic analyses as preliminary pilot tests (results not shown) revealed generally lower concentrations of DNA and higher levels of DNA fragmentation compared to freshly-processed specimens.

Based on DNA concentrations (above or below 7.0 ng/µl) and DNA fragmentation (> 350 bp or highly fragmented defined as < 350 bp), samples were submitted to Berry Genomics (n = 563) for standard library preparation or to Novogene (n = 81) for low input DNA library preparation, respectively. All samples were sequenced at 10x coverage on an Illumina NovaSeq platform (150 PE reads, 6 Gb raw data output/sample). Quality parameters of extracted DNA and WGS data of specimens originating from five insect genera and more than 70 different species were collected (see Table 3 and Suppl. material 2).

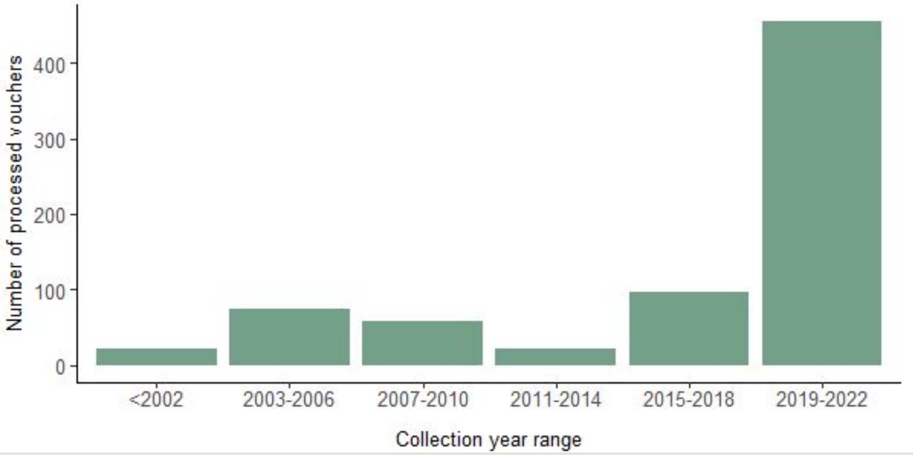

Figure 2. doi

Age distribution of processed specimens of fruit flies and hoverflies.

Table 3.

Number of processed collection vouchers from three Tephritidae and two Syrphidae genera.

Collection	**Genus**	**Number of specimens**
Tephritidae	*Bactrocera*	16
Tephritidae	*Dacus*	197
Tephritidae	*Ceratitis*	411
Syrphidae	*Eristalinus*	83
Syrphidae	*Melanostoma*	25

Results and Discussion

Overall, our DNA extractions provided detectable DNA in 720 specimens out of 732 (98.4%) while WGS data could be obtained for 644 museum vouchers (88.0%). The different DNA extraction methods gave broadly similar yields, albeit with a somewhat lower recovery of DNA from whole-body extractions using the MinElute kit. Overall, there was a heterogeneous recovery of DNA yields across specimens (Fig. 3), with values ranging from 57.8 to 153.0 ng for whole bodies. As expected, due to the lower amount of tissue, DNA amounts recovered from legs was comparably lower and ranged from 1.3 to 22.0 ng. Based on minimising costs, we adopted the kit with the lowest price (DNeasy Blood and Tissue Kit) for routine processing of vouchers from the target insect collections.

Our results show a general trend of decreasing recovery of DNA from older specimens compared to younger specimens (Fig. 4a). In contrast, our assessment of DNA quality as estimated by fragment lengths of the DNA extracts and Phred score (Q > 30) of raw sequences lacks any clear temporal trend, with degraded DNA with short fragment lengths and low quality reads recovered across the range of specimen ages (Fig. 4b and c).

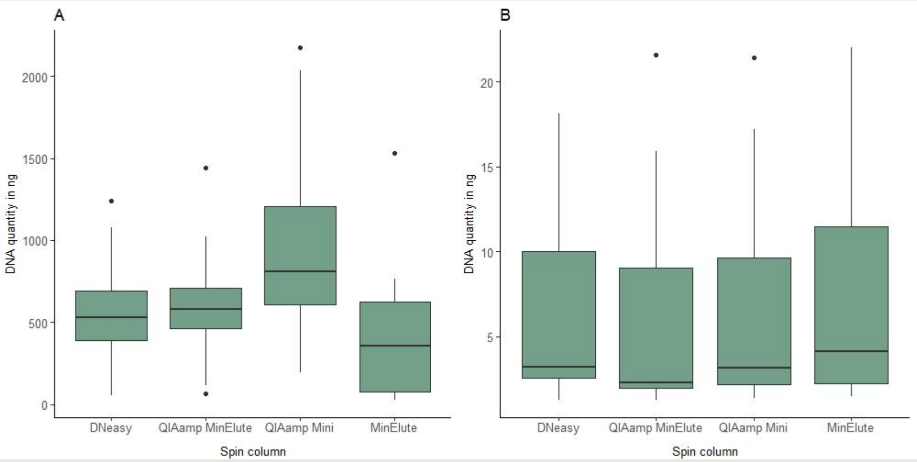

Figure 3. doi

Boxplots of DNA yields from replicated elutions of Tephritidae and Syrphidae of **A** whole body digestions and **B** leg digestions per DNA extraction method (Qiagen, DNeasy Blood and Tissue Kit, cat. 69506; QIAamp DNA Mini, cat. 51304; QIAamp DNA Micro, cat. 56304; MinElute PCR Purification, cat. 28004).

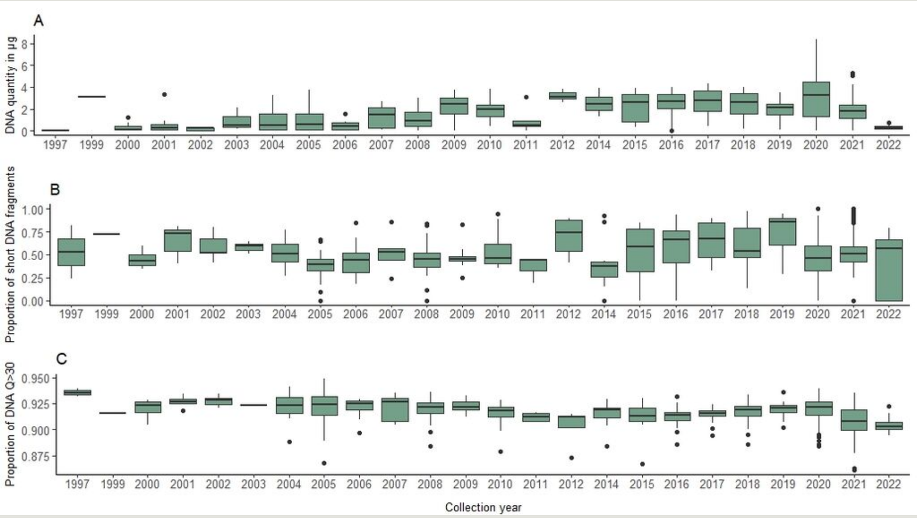

Figure 4. doi

Boxplots per collection year for Tephritidae and Syrphidae specimens extracted with the DNeasy Blood and Tissue Kit (Qiagen): **A** DNA quantities (calculated from concentration measured with Qubit 4.0); **B** the proportion of DNA fragments between 35 and 350 bp (measured with Fragment Analyzer (DNF-930 dsDNA Reagent kit)); **C** proportion of sequenced reads with Q > 30.

Sub-optimal or low-quality DNA from museum specimens is often not directly suitable for genetic/genomic analyses (Besnard et al. 2016). However, our results suggest 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 insect vouchers collected over the past two decades. The majority of samples processed in this analysis were of material stored in ethanol. The samples that were pinned at room temperature (n = 14) or stored as dried DNA at room temperature (n = 67) showed similar DNA quantity and quality results as the DNA from specimens stored in ethanol.

The DNA of these diverse and heterogeneously collected samples, even if generally suboptimal in terms of concentration, fragmentation and contamination, still allowed recovery of substantial amounts of quality reads (Q > 30) of potential use for genomic research. This general approach needs to be complemented with more specialised and time/cost-demanding procedures for highly-degraded DNA from older specimens. A two-step approach, including the use of commercial kits and methods outlined here allows for rapid screening of younger specimens and reserving the more intensive protocols (also including aDNA methodologies) for older specimens represents a pragmatic, cost-effective route to the routine genotyping of our insect collections.

Main findings and recommendations:

- The DNeasy Blood and Tissue Kit (Qiagen) provided a cost-effective method of extracting DNA from museum specimens aged one to 25 years;
- These recently-collected samples, although containing fragmented DNA, represent a tractable tissue source for large-scale sequencing projects;
- For older material, the use of low input library preparation for highly-fragmented and low concentration DNA extracts is recommended.

Data availability: The data and meta-data from this case study are documented in Suppl. material 2.

Case study 3: Genome skimming as a tool to recover whole plastid genomes from threatened Central African timber species

Introduction

Worldwide, multiple tree species used for timber production are under severe threat (Fig. 5). Despite a restriction on logging concessions and the improvement of forestry laws, recent studies show that, for example, in the Democratic Republic of Congo, illegal tree logging represents over 75% of the annual industrial timber production (Nellemann 2012). DNA-based identification tools can support investigations into illegal trade, but depend upon an accurate genetic reference database to identify and trace the provenance of logged trees. In this regard, herbarium collections are an excellent source to generate such genetic reference databases, especially in areas where field expeditions are not feasible anymore due to political instability or increased inaccessibility. Here, we demonstrate the usefulness of genome skimming by shotgun sequencing to mine herbarium specimens for the assembly of their plastomes to support DNA-based identification of trade timber species.

Figure 5. doi

Transportation of large tree trunks from the forest to enter the international timber trade.

The quality and quantity of DNA in herbarium specimens is strongly reliant on collection and storage conditions and, in general, herbarium DNA can be highly fragmented (< 150 bp) and only available in very low amounts (< 5 ng/µl). Interestingly, the techniques optimised for historical herbarium specimens can also be applied to heartwood specimens (i.e. old degenerated material) of processed wood. By jointly analysing herbarium material and silica-dried leaf samples, a clear comparison can be made on the feasibility of historical material for genome skimming purposes, with the aim of yielding full plastid genomes of selected species that are under strong pressure due to illegal logging activities in Central Africa.

Materials and Methods

In order to obtain plastomes of the most important timber species from equatorial African tree species, we collected leaf tissue samples (2 cm^2; ca. 10 mg) from 12 herbarium specimens and 20 silica samples via various herbaria (BR, BRLU & L; Suppl. material 3).

Tree species were selected, based on the following criteria: providing highly valuable timber, becoming potentially important for national and international timber trade or for being reported in agreements on global biodiversity conservation (e.g. CITES, IUCN). In the case of herbarium samples, there was a careful selection based on prior knowledge about the specimens. We avoided material that was likely to have been: (a) dried with alcohol, (b) treated with conservatives posterior to collection (e.g. $HgCl_2$) or (c) collected in remote areas where it was difficult to properly dry the specimens in the field. When sampling from herbarium specimens, we aimed to: (d) collect the greenest leaf tissue, (e) avoid the central leaf vein and (f) avoid leaves with potential markings of insect herbivory.

DNA extraction and library preparation

Total genomic DNA of both silica and herbarium material was extracted using a combined and modified version of the CTAB protocol (Doyle and Doyle 1987) in which an initial washing step with 0.35 M d-sorbitol was included. The lysis buffer contained 2% CTAB and 2% PVP-40. A chloroform-isoamyl alcohol (24/1 v/v) extraction was carried out twice. After a cold isopropanol precipitation and subsequent centrifugation, the pellet was washed with 70% ethanol and air-dried. The DNA pellet was eluted with 1X TE buffer. All herbarium specimen DNA extractions were carried out under a laminar flow hood, in which positive air pressure and UV disinfection was present.

The purity of the resulting DNA was measured under the absorbance ratios OD 260/280 and OD 260/230 using a NanoDrop 2000 (Thermo Fisher Scientific, US). DNA concentration (ng/µl) and fragment size distribution were measured by capillary electrophoresis using a Fragment Analyzer (Agilent, US). Library preparation of the silica-dried leaf material was initiated via an enzymatic DNA fragmentation step with the aim to retain DNA fragments with a size between 200 and 450 bp after which an end repair step took place. This step was conducted using the NEBnext UltraTM II FS DNA Library Prep Kit for Illumina (New England Biolabs, US). Due to the presence of already degraded DNA in the herbarium specimens, the enzymatic DNA fragmentation step was not carried out for the herbarium material. Adapter ligation was conducted using the NEB-next Adaptor kit for Illumina and U-excision was carried out using the USER Enzyme kit (New England Biolabs, US). Size selection (320–470 bp) was conducted following the SPRIselect protocol (Beckman Coulter, US). With the NEBNext Ultra II Q5 Master Mix, adaptor-ligated DNA was indexed, then PCR-enriched with the NEBNext Multiplex Oligos for Illumina (New England Biolabs, US). For the latter, the following thermocycler reactions were used: Initial denaturation at 98°C for 30 s, 3–4 cycles of denaturation at 98°C, each for 10 s as well as an annealing/extension at 65°C for 75 s and a final extension phase at 65°C for 5 min. In the last step of the library preparation, a DNA-library purification was conducted using SPRIselect (Beckman Coulter, US). The final fragment size distribution and molarity (nM) were examined with a Fragment Analyzer (Agilent, US). Indexed libraries were subsequently pooled (on average 25 samples per lane) in equimolar ratios. Sequencing of the DNA libraries (low coverage paired-end; 10X, 150 bp) was done on a HiSeq 3000, a HiSeq 4000 and NovaSeq 6000 (Illumina, US). At the time of analysis, between autumn 2019 and spring 2021, library preparation and sequencing costs were estimated at a total of €45 - 50 per sample.

Data analysis

The quality of the raw reads was evaluated with FastQC (Andrews 2010). Using the GetOrganelle pipeline, plastomes were de-novo assembled (Jin et al. 2020) under *k-mer* values set as "-k 21, 45, 65, 85, 105" for all species and 15 extension rounds. The pipeline was initiated by recruiting targeted plastid-like reads as applied in Bowtie 2 (Langmead and Salzberg 2012). During the assembly process, reads were trimmed and contigs reconstructed with SPAdes 3.13 (Bankevich et al. 2012). In addition, plastid-like contigs were filtered by comparing them against the BLAST nucleotide database following the NCBI Blast+ tool (Camacho et al. 2009). In the next step, reconstructed plastomes were aligned against a reference genome with MAFFT v.7 (Katoh et al. 2002), thereby aiming for the most closely-related taxon for comparison that could be found on GenBank. In the case of unsuccessfully assembled plastome regions using the GetOrganelle pipeline, raw reads were then mapped to target regions of closely-related species using Bowtie 2 (*Guibourtia pellegriniana* and *G. tessmannii*). Applying the web-based software CpGAVAS2 (Shi et al. 2019), annotation was conducted, after which the annotation results were endorsed with Geneious Prime (Kearse et al. 2012) by comparing them with a reference plastome derived from GenBank.

Results and Discussion

Amongst the 16 herbarium specimens, DNA yields varied between 220 and 430 ng/µl, whereas DNA yields of silica-dried samples varied between 210 and 850 ng/µl (based on starting leaf tissue samples of 2 cm^2). All taxa investigated yielded sufficient DNA and were used for library preparation and sequencing. For the herbarium specimens, between 200,000 and 5.2 million high-quality paired-end reads were produced, whereas for the silica-dried samples, this amount varied between 1.6 million and 6.4 million reads. Over 4 million high-quality paired-end reads were retrieved for only 19% of the herbarium specimens, whereas for the silica-dried samples, there was a ca. 50/50 split of accessions above or below 4 million high-quality paired-end reads. Genome coverage depth was at least 50x for all specimens with fully assembled plastid genomes. Even though a very small number of reads were retrieved from some specimens, it was possible to generate complete plastomes for the majority of the herbarium specimens (Mascarello et al. 2021). The complete plastomes always consisted of a small circular sequence partitioned in four main structures that are typical for land plants: a large single copy region (LSC) and a small single copy region (SSC), disconnected from each other by two inverted repeats (IRa and IRb). There was only one specimen where the number of reads was below 1 million from which a full plastome could not be obtained. Moreover, for this accession, the percentage of duplicate reads was 5%, whereas for all other accessions, the percentage of duplicate reads varied between 9 and 13%. The quality of the plastome sequence was checked by translating all gene regions. No stop codons (indicative of sequencing errors) were observed along the assembled contigs. Using this approach, a genetic reference database of threatened African trees has been developed as a tool against illegal logging, as well as an optimised DNA isolation protocol to obtain sufficient DNA via the Genome Skimming by Shotgun Sequencing method (Cronn et al. 2008, Mascarello et al. 2021).

The results obtained in this case study corroborate those of some recently-published studies on the use of genome skimming for the retrieval of full plastomes of land plants (Bakker et al. 2016, Zeng et al. 2018, Alsos et al. 2020, Nevill et al. 2020). Each of those studies indicates the scalable potential of genome skimming to obtain plastome sequences from herbarium specimens. Using this approach, a high level of success has been achieved across a range of ages of herbarium specimens (Alsos et al. 2020, Nevill et al. 2020) and even with small amounts of tissue, an effective plastome assembly can be generated (Bakker et al. 2016). This collective body of studies shows that genome skimming represents an inexpensive, pragmatic approach for recovery of plastome sequences that can be applied to small amounts of herbarium material.

Main findings and recommendations:

- Genome skimming of herbarium specimens has shown high success rates across multiple independent studies;
- Despite the often lower number of reads retrieved from herbarium specimens compared to fresh tissue, it is becoming increasingly routine to generate complete (or almost complete) plastomes from herbarium material using genome skimming;
- Since one of the most important steps in the genome skimming protocol is to downsize the DNA fragment length, the often highly-degraded DNA of herbarium specimens allows the sonication step to be bypassed in the library preparation protocol.

Data availability: The data in this case study are available under the following GenBank numbers: MZ274087, MZ274092, MZ274094-MZ274096, MZ274099, MZ274102-MZ274107, MZ274110, MZ274113, MZ274116-MZ274122, MZ274124, MZ274127-MZ274129, MZ274132, MZ274135, MZ274137, MZ274143, MZ274145, MZ274147, MZ274148 (see Mascarello et al. (2021)).

Case study 4: Comparing hybridisation capture derived sequences from herbarium specimens with data from living material of the same genetic individuals

Introduction

Herbarium collections worldwide contain an estimated 350 million specimens collected over the last approximately 400 years (Besnard et al. 2018). Advances in DNA extraction and sequencing technologies are making herbarium specimens increasingly accessible to retrospective genomic analyses. Sequencing approaches include shotgun metagenomics (Bieker et al. 2020), Whole Genome Sequencing (WGS) (Yoshida et al. 2013), genome skimming (Bakker et al. 2016, Zeng et al. 2018, Nevill et al. 2020, Case study 3) and hybridisation capture (Hart et al. 2016, Sánchez Barreiro et al. 2017, Gutaker et al. 2019).

The preservation and quality of DNA in herbarium material are highly variable. It has been suggested that DNA decays at a faster rate in plant remains compared to animals (Allentoft et al. 2012, Weiß et al. 2016) and the techniques used for herbarium sheet preparation and the storage conditions of specimens have been shown to affect DNA recovery and fragmentation rates (Särkinen et al. 2012, Forrest et al. 2019). Studies comparing recently prepared and older herbarium specimens do not reach a consensus on whether DNA fragmentation and damage happen mainly at specimen preparation (e.g. Staats et al. (2011)) or accumulate gradually over time (e.g. Weiß et al. (2016)). These discrepancies are likely the result of different preparation techniques, ranging from gentle drying in non-acidic paper to high heat and chemical treatments. For this reason, for herbarium specimens, it is common to see both laboratory protocols aimed at recovering and sequencing low-concentration, fragmented DNA (e.g. Latorre et al. (2020)) and protocols commonly used for higher-quality DNA sources (e.g. Forrest et al. (2019)).

In this study, we sampled specimens from the herbarium at the Royal Botanic Garden Edinburgh (RBGE) that were collected 12-50 years ago from cultivated individuals of *Rhododendron javanicum*. These cultivated individuals are still present as live plants in the living collection at RBGE and allowed us to assess the reliability of sequences recovered from herbarium material compared to freshly-collected samples from the same genetic individuals. The chosen sequencing approach was hybridisation capture (also known as target capture or target DNA enrichment) which is an effective sequencing approach for studies utilising degraded DNA sources because it enables recovery of sequence data from low concentrations of endogenous DNA (Carpenter et al. 2013).

Materials and Methods

Samples

Twelve RBGE herbarium vouchers (dated 1972-2010) of various sub-species of cultivated *Rhododendron javanicum* were sampled along with fresh leaf material from 10 living individuals growing in the RBGE glasshouses, from which the herbarium vouchers were generated. Two of the living individuals were represented by two separate vouchers, collected one or ten years apart. Herbarium samples and their corresponding living samples are listed in Table 4. Additional sample information is provided in Suppl. material 4.

DNA extraction and library preparation - herbarium samples

DNA from herbarium specimens was extracted as described in Latorre et al. (2020), using the Basic Protocol 1, following standard anti-contamination precautions (Gilbert et al. 2005, Llamas et al. 2017), including parallel non-template controls. DNA fragment size distribution for extracts was inspected with the gDNA Kit on the Agilent Femto Pulse System. Sequencing library preparation protocol was selected based on DNA fragment size (Table 4).

Table 4.

Details of *Rhododendron* herbarium samples used in this study including collection and accession numbers, as well as library protocols used. RHD002 and RHD007 herbarium specimens relate to the same single individual in the living collection, as do RHD016 and RHD018, respectively. Two samples (RHD011 and RHD018) had sequencing libraries prepared using two different protocols. Fresh samples from the living collection were also collected for all individuals. DNA fragment size distribution: size as stated, except bimodal which means one peak of < 1000 bp and one peak of approximately 1-20 kbp. ssDNA = single-stranded DNA library, NEB = NEBNext Ultra II library with sonicated DNA. *All sequencing libraries for the *living collection* were prepared using NEBNext Ultra II kits with sonicated DNA.

Sample	**Species**	**Subspecies**	**RBGE herbarium collection number**	**Specimen date**	**RBGE living collection accession number**	**DNA fragment size distribution**	**Library protocol(s) for herbarium samples***
RHD002	*Rhododendron javanicum*	*kinabaluense*	E00421003	2010	19801291A	bimodal	NEB
RHD003	*R. javanicum*	*moultonii*	E00294943	2009	20110223A	bimodal	NEB
RHD005	*R. javanicum*		E00328126	2009	19672627A	100-1000 bp	NEB
RHD006	*R. javanicum*	*brookeanum*	E00328133	2009	19801298C	bimodal	NEB
RHD007	*R. javanicum*	*kinabaluense*	E00328548	2009	19801291A	bimodal	NEB
RHD008	*R. javanicum*	*palawanense*	E00294512	2008	19922762B	bimodal	NEB
RHD009	*R. javanicum*	*cladotrichum*	E00294755	2007	19913084A	bimodal	NEB
RHD011	*R. javanicum*	*palawanense*	E00954297	1998	19922772	bimodal	ssDNA, NEB
RHD013	*R. javanicum*	*javanicum*	E00954260	1990	19730741	< 500 bp	ssDNA
RHD016	*R. javanicum*	*javanicum*	E01016321	1982	19680840	< 500 bp	ssDNA
RHD017	*R. javanicum*	*kinabaluense*	E01016323	1981	19690955	< 500 bp	ssDNA
RHD018	*R. javanicum*	*javanicum*	E01016322	1972	19680840	< 500 bp (+tail)	ssDNA, NEB

DNA extracts with fragments shorter than 500 bp (n = 5) were converted into single-stranded DNA (ssDNA) libraries following Kapp et al. (2021) with tier four adapter dilutions and unique dual indexes. All steps up to indexing PCR reactions were carried out in dedicated ancient DNA facilities at the University of Oslo, Norway. Sequencing library quality and concentration were assessed by qPCR following Meyer and Kircher (2010) and with the Ultra Sensitivity NGS Kit on the Agilent Femto Pulse System. Libraries were then re-amplified in three 25 µl reactions with Herculase II Fusion DNA polymerase (Agilent). One sample (RHD011) with longer DNA fragments was also included in this library preparation batch.

Of the remaining samples (n = 9), we recovered less-fragmented DNA, including samples with a bimodal DNA fragment size distribution (n = 7), with one peak of fragments shorter than 1000 bp and a second peak of approximately 1-20 kbp; the other samples included one sample (RHD005) with DNA fragments of 100-1000 bp and one (RHD018) with mostly short fragments, but with a tail of longer fragments. Aliquots of these extracts were subjected to 8-12 sonication cycles of 30 s on, 90 s off, using a Diagenode Bioruptor sonicator, for a target fragment size of 200-400 bp. Libraries were generated with the NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (New England Biolabs) and indexed with NEBNext® Multiplex Oligos for Illumina® (Unique Dual Index Primer Pairs). Because these libraries were produced in non-dedicated facilities where fresh plant material is regularly processed — posing a risk for contamination — the following anti-contamination precautions were taken: pre-amplification steps were carried out inside a dedicated laminar flow hood in a pre-PCR room with dedicated reagents and consumables and negative non-template controls were included.

DNA extraction and library preparation - living collection samples

Approximately 150 mg of leaf material was harvested into 7.6 ml FluidX tubes and placed immediately into liquid nitrogen. DNA was extracted using a protocol developed for extracting high molecular weight DNA (Nishii et al. 2023). This protocol, which includes a sorbitol wash prior to using the Qiagen Genomic Tip Kit, was used due to the high quantity of secondary metabolic compounds present in *Rhododendron*. The DNA extracts were sonicated for 7-11 cycles of 30 s on, 90 s off, using a Diagenode Bioruptor sonicator, for a target fragment size of 200-400 bp. Libraries were generated with the NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (New England Biolabs) and indexed with NEBNext® Multiplex Oligos for Illumina® (Unique Dual Index Primer Pairs). These libraries were generated in non-dedicated facilities with pre-amplification steps carried out inside a dedicated laminar flow hood with dedicated reagents and consumables.

Hybridisation capture and sequencing

Hybridisation capture was performed on all libraries. The assay was designed using two published *Rhododendron* genomes from NCBI: *R. delavayi* (Zhang et al. 2017) and *R. williamsianum* (Soza et al. 2019) and a transcriptome from the mature leaf of *R. scopulorum* from the 1000 Plants (1KP) project (Matasci et al. 2014). The bait set contains 492 target loci, including 298 loci orthologous to the Angiosperm353 universal probe set (Johnson et al. 2019). The remaining 194 loci were picked from genes related to cold tolerance, flowering pathway, key developmental regulators of meristem function, organ development and trichome development. Baits were synthesised by MyBaits (Arbor Biosciences) with 3X bait tiling to be optimal for degraded DNA.

Libraries were pooled according to material and library construction protocol. The samples were processed with a wider set of samples than are presented here, such that each pool contained 10-14 libraries. *Negative controls* were pooled separately. Hybridisation capture was performed following the MyBaits (Arbor Biosciences) protocol v.5.02 with the high sensitivity version and the hybridisation and wash temperatures set to 63°C for herbarium samples (the second round of enrichment was omitted) and with the standard version and hybridisation and wash temperatures set to 65°C for living samples. Pools were re-amplified post-capture in two 50 μl reactions with Herculase II Fusion DNA polymerase (Agilent) for 14 cycles. Captured libraries for living and herbarium samples were sequenced on separate Illumina MiSeq lanes with no index repetition, with 150 bp PE v.2 runs (4.5 - 5.0 Gb) at the University of Exeter sequencing facilities. The target region of the bait set is 621,078 bp, which translates to an average coverage of 720X/sample for the 10 living samples and 600X for the 12 herbarium samples.

Data analysis

Herbarium reads were processed with the PALEOMIX v.1.3.7 BAM pipeline (Schubert et al. 2014). Paired-end reads were trimmed, filtered, and collapsed with AdapterRemoval v. 2.3.3 (Lindgreen 2012), discarding reads shorter than 25 bp. Collapsed reads were aligned to the target loci used for probe design with BWA v.0.7.17 (Li and Durbin 2009), using the backtrack algorithm with disabled seeding and a minimum quality score of 25. mapDamage v.2.2.1 (Jónsson et al. 2013) was used to assess aDNA deamination patterns and rescale BAM file quality scores. Living collection reads were processed as described for herbarium reads without read collapsing and retaining reads longer than 50 bp. The BWA MEM algorithm was used for read alignments to the same references.

Quantity and quality of the SNPs called for the herbarium samples were assessed by comparison to the sequence from their respective paired living sample. First, a new reference for each individual was generated using sequence data from only the living sample of that individual. BAM files from the initial run of PALEOMIX (above) were filtered using strict settings on bcftools v.1.16 (filter SNPs by QUAL > 160 and DP > 10) and consensus fasta files were generated to be used as a new reference (Forrest et al. 2019). The new reference was used to run the PALEOMIX BAM pipeline for a second round for the same living and their respective herbarium sample pairs, this time using the individual new references rather than the original target sequences. New VCF files were generated from the output BAM files and bcftools stats was used to compare SNPs called from the living and from the herbarium material. We identified shared SNPs in living and herbarium samples, but not present in the new reference (likely heterozygous sites) and those exclusive to the herbarium samples (likely erroneous SNPs). The code used to analyse the data and make figures is available at: https://github.com/rbgedinburgh/dna_sequencing_herbaria.

Results and Discussion

Evaluating sequencing library preparation for herbarium material and contamination control

Without any prior assumption of DNA fragmentation rates in the herbarium samples processed in this study, our approach consisted of isolating DNA in dedicated clean facilities. Following an assessment of DNA fragment size, we decided to separate samples into two categories. Fragmented DNA extracts were treated, following aDNA protocols in dedicated facilities, whereas extracts containing longer DNA fragments were taken to non-dedicated facilities for DNA shearing and library preparation using commercially available kits. We assessed coverage of targeted loci (Fig. 6A) and library complexity —using read clonality — (Fig. 6B) by mapping reads to the target loci used for probe design. As expected, we obtained higher coverage of targeted loci from freshly-collected samples for similar amounts of sequencing effort. For herbarium samples, libraries generated from sheared DNA using the NEB kit had higher complexity and higher coverage of targeted loci than ssDNA libraries that were generated from highly-degraded DNA. This is to be expected given the difference in quality of input DNA. Detailed mapping statistics are available in Suppl. material 5.

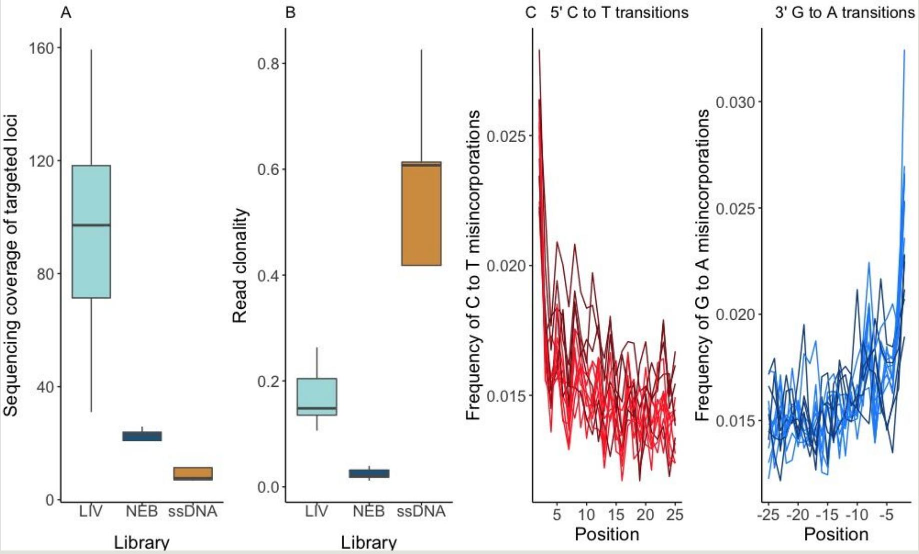

Figure 6. doi

Sequencing coverage of targeted loci and library complexity for *Rhododendron* specimens. **A** Coverage of targeted nuclear loci. **B** Proportion of PCR duplicates. LIV = libraries generated from living collection samples, ssDNA = single-stranded DNA libraries made from degraded herbarium DNA, NEB = double-stranded DNA libraries made from sheared herbarium DNA using a commercial kit. **C** DNA deamination patterns of read data obtained from NEB (red, blue) and ssDNA (dark red, dark blue) herbarium libraries with mapDamage v.2.2.1. First base was removed for visualisation.

We did not detect any amplification products in the negative controls of libraries generated in the non-dedicated facilities following anti-contamination measures, indicating that it was possible to process herbarium DNA extracts of sufficient DNA concentration and fragment size under these conditions and with the necessary precautions. In addition to the ssDNA library construction protocol (Kapp et al. 2021), we also tested a dsDNA protocol optimised for aDNA (Meyer and Kircher 2010, Kircher et al. 2012) (data not shown), but we observed high levels of sequence clonality, possibly caused by PCR inhibition. DNA isolation and sequencing library preparation for plant material can be complicated by secondary compounds, such as polysaccharides and polyphenols that can bind to and co-precipitate with DNA, resulting in PCR-inhibition (Souza et al. 2012). *Rhododendron* is rich in secondary metabolic compounds (which also led to difficulties in extracting DNA from fresh material) and it is possible that the initial DNA denaturation step in the ssDNA library preparation protocol had a beneficial effect on breaking crosslinks between DNA and secondary compounds (compared to the dsDNA protocol). We only tested a small number of samples, but the efficacy of this comparatively fast and cheap ssDNA protocol is promising and further testing on short degraded DNA isolated from herbarium material would be worthwhile.

Finally, we observed mild deamination patterns in reads recovered from herbarium material (Fig. 6C) compatible with historic DNA damage (Jónsson et al. 2013), although the magnitude of this was very small (ca. 3% first base misincorporation) compared to levels often observed in older material. Interestingly, we observed similar deamination patterns in libraries generated with the ssDNA protocol and the NEB kit, indicating that despite DNA shearing, and the NEB library preparation including a USER enzyme hairpin loop adaptor cleavage step, enough base deaminations at DNA overhangs were retained to show evidence of post-mortem DNA damage (Jónsson et al. 2013).

Assessing reliability of SNPs recovered from herbarium material

We took advantage of cultivated plants present in the RBGE living collection, from which herbarium vouchers were created 12-50 years ago, to investigate whether sequences recovered from the herbarium samples were an accurate biological replicate of the living material, or if low starting templates and base modifications, both features that accumulate in degrading DNA over time, resulted in erroneously-called bases (Briggs et al. 2007). Using only sequences recovered from living material, we assembled a strict consensus sequence for each individual. These were used as a new reference for mapping and SNP calling. We assigned SNPs as being exclusive to living samples, exclusive to herbarium samples or shared between a living-herbarium pair of the same individual. SNPs exclusive to living samples might be caused by ambiguous calls at heterozygous sites, while SNPs exclusive to herbarium samples can be interpreted as erroneous SNPs, likely due to low SNP quality, low coverage or base modifications in degraded DNA. In contrast, shared SNPs between living and herbarium tissue can be interpreted as true.

We typically observed 75-108 SNPs per individual, of which 45-87 were shared between living and herbarium samples (Fig. 7A) and 12-36 that were found in herbarium specimens only. We did not observe a correlation between specimen age and proportion of these likely erroneous SNPs. We also inspected the quality and sequencing depth of SNPs and found that SNPs exclusive to herbarium samples were of much lower quality than those present in both herbarium and living material (Fig. 7B and C). The quality and depth of these erroneous SNPs unique to herbarium specimens are, however, above standard SNP-filtering thresholds. In our study, the use of more stringent filtering criteria for herbarium SNPs is required to give a better representation of 'true' sequence variants (e.g. those also recovered from non-degraded tissues).

Two samples (represented by three libraries) showed a noticeable spike in SNP abundance compared to all others. Both of these samples are from the same sub-species *Rhododendron javanicum* ssp. *palawanense* (RHD008 and RHD011) and retrospective analyses of their morphology suggest they may be of hybrid origin. It is possible that the observed spike in SNP abundance is due to these specimens having higher levels of heterozygosity due to hybridity. With a greater number of variable sites, there is an associated increased possibility of detecting both genuine (shared) SNPs with respect to the reference, as well as a corresponding increase in erroneous SNPs due to poor coverage of these sites in herbarium material.

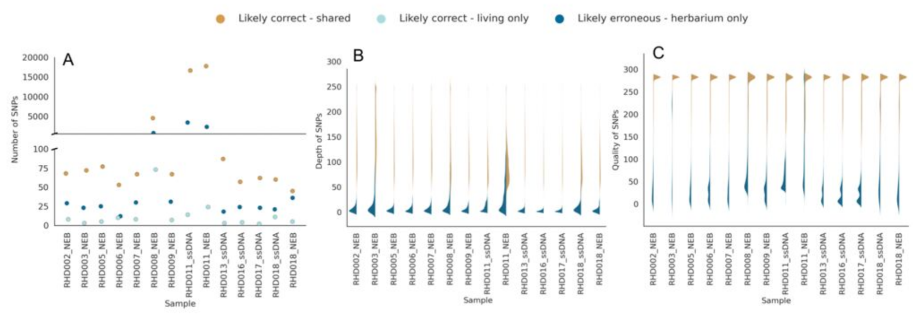

Figure 7. doi

Comparison of SNPs recovered from herbarium and living collection samples of the same individuals of *Rhododendron* species. **A** Number of SNPs called, categorised into exclusive to living samples (light blue, likely caused by ambiguous calls at heterozygous sites), exclusive to herbarium samples (dark blue, likely caused by sequencing errors due to degraded DNA) or shared (yellow). **B** Depth and **C** quality of shared and herbarium-exclusive SNPs. ssDNA = single-stranded DNA libraries made from degraded herbarium DNA, NEB = double-stranded DNA libraries made from sheared herbarium DNA using a commercial kit.

Implications for sequencing herbarium specimens

Multiple studies have now been undertaken exploring the potential of hybridisation capture for the recovery of sequence data from herbarium specimens. These have included exploratory studies assessing the feasibility of the approach for recovering sequence data from plant specimens with a range of different ages (Hart et al. 2016), studies evaluating the impacts of different treatment methods on sequencing success (Brewer et al. 2019, Forrest et al. 2019) and those exploring the practicalities of scaling hybridisation capture in plants, including how the characteristics of specimen origin and condition influence sequence recovery (Folk et al. 2021, Kates et al. 2021). Collectively, these and other studies have provided clear evidence that the recovery of large amounts of nuclear sequence data is feasible for herbarium specimens with a wide range of ages and across different taxonomic groups. In the current case study, we have shown that erroneous base-calls can be made due to low starting templates and modified bases in herbarium specimens. However, with stringent filtering for quality and depth, these erroneous SNPs can be excluded, such that the remaining SNPs represent a more accurate reflection of the individual's genotype.

Main findings and recommendations:

- Hybridisation capture is now well established as a method for recovery of large amounts of nuclear sequence data from herbarium specimens and the approach works well in accommodating the complexity of plant genomes;
- Studies recovering DNA from herbarium specimens should take place in dedicated clean or low-copy facilities. Once DNA fragment length distribution is known, sequencing library preparation can take place according to DNA size;
- Library preparation from highly-fragmented DNA should take place in dedicated clean or low-copy facilities. We found the ssDNA protocol by Kapp et al. (2021) to be fast and efficient for this purpose in the current study and the simplicity of this approach warrants further trialling to see if these results are generally applicable;
- DNA extracts that show a bimodal fragment size distribution with the majority of fragments > 1 kb can be sheared, prior to library preparation with a commercially-available kit. If this takes place in non-dedicated facilities, we recommend the following contamination-limiting precautions:
  - Physical separation of pre- and post-PCR laboratories;
  - Dedicated laminar flow hood for all pre-PCR steps (to be regularly decontaminated);
  - Dedicated reagents and consumables;
  - Inclusion of non-template negative controls;
- Distribution of SNP quality and coverage should be inspected for a better-informed decision on filtering parameters. Stringent quality filtering of SNPs can provide high confidence in genotype calls even from herbarium material.

Data availability:

The raw sequencing data has been deposited at the European Nucleotide Archive (ENA) under project PRJEB61704 with accession IDs ERS15567903 - ERS15567924.

Case study 5: Selecting samples with the greatest likelihood of success for reduced-representation sequencing from museum collections

Introduction

Reduced-representation sequencing (RRS) using restriction digests followed by fragment sequencing is a cost-effective route for generating thousands of genetic markers (Davey and Blaxter 2010, Davey et al. 2011, Puritz et al. 2014). Although this type of approach has proved very effective when working with high-quality DNA (Nadeau et al. 2014, Van Belleghem et al. 2018), its implementation in museum studies has been hampered by the unpredictable outcomes due to DNA degradation of museum specimens (Graham et al. 2015, Souza et al. 2017, Lang et al. 2020). DNA degradation at restriction sites causes failure or bias in RRS due to inefficient or failed restriction digests, while random shearing lowers the number of fragments being flanked by both restriction sites and therefore prevents 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, the number of variable sites and the percentage of identical RADtags retained (Graham et al. 2015). These difficulties have dissuaded scientists from using RRS as a tool to obtain museum population-level data. However, when large collections are available, a careful screening assessment prior to library preparation could aid in the selection of samples that are most likely to yield successful results. Therefore, we here assess: (i) to what extent DNA degradation affects the success rate of RRS in a long-term time series of avian museum specimens and (ii) whether we can predict *a priori* the success rate of RRS on museum samples using easy-to-obtain DNA quality metrics.

Materials and Methods

Sampling

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, mainly from the 1930s (1929-1943, n = 15) and mainly from the 1970s (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 minimise voucher damage and liver or breast muscle tissue of the contemporary specimens.

DNA extraction, library preparation and SNP calling

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 focused on 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 (18 µl, 12 µl or 6 µl of DNA when concentrations were respectively lower than 20 ng/µl, between 20 and 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 reverse primer to obtain dual-indexed ddRAD libraries, which were subsequently paired-end sequenced on an Illumina Novaseq6000 platform. Raw data were demultiplexed using the process_radtags module in Stacks v.2.50 (Catchen et al. 2011). Trimmomatic v.0.39 (Bolger et al. 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 v.1.11 (Li et al. 2009). SNPs were subsequently called using GATK's HaplotypeCaller tool (McKenna et al. 2010).

Contamination assessment

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 dataset of 31,012 SNPs covering 10,349 RADtags and 0.3% of the genome. These reference SNPs were then subsequently extracted from all individuals, for example, both historical as well as contemporary specimens, to limit the erroneous inclusion of exogenous DNA sequences from historical samples. As the SNP discovery protocol is exclusively applied on recent samples, this could, however, result in a SNP ascertainment bias and concomitant underestimation of genetic diversity in historical populations or erroneous measures of genetic differentiation (Lachance and Tishkoff 2013). To eliminate such bias, one should identify a sufficient number of high-quality historical samples with minimal missing data and repopulate the SNP discovery pipeline with this extended dataset.

Statistical analysis

We ran a one-way ANOVA to test for differences in mean number of missing SNPs amongst the three time periods and allowed for period-specific variances to account for heteroscedasticity using the R package 'nlme' (Pinheiro et al. 2022). To predict the success rate of ddRAD in museum samples, we applied generalised additive models (GAM) to relate the 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 area under the curve in four distinct bins, i.e. bins that contain fragments ranging from, respectively, 35-200 bp, 200-400 bp, 400-700 bp or 700-10380 bp.

Results and Discussion

Mean missing data per individual differed significantly between time periods (chi-square = 62.56, $p < 0.001$) (Fig. 8). The mean percentage of missing SNPs was 2.6% for recent specimens, 43.4% for specimens sampled around the 1970s and 85.4% for specimens originating from around the 1930s. The variance in missing data varied significantly between time periods (Breusch-Pagan test, chi-square = 52.1, $p < 0.001$). Recent samples showed consistently few missing SNPs, while the success rate in samples from the 1930s varied slightly more. In contrast, samples from the 1970s 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.

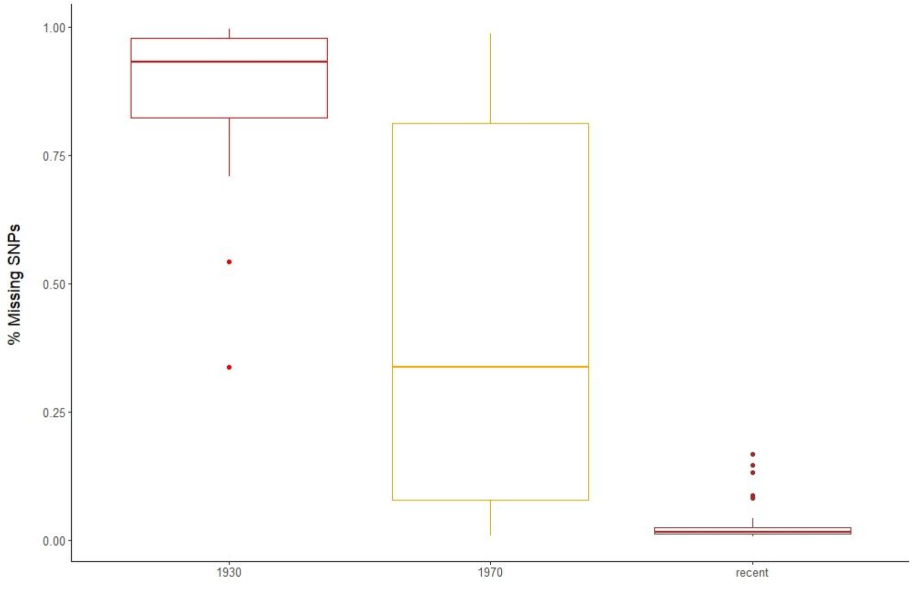

Figure 8. doi

Percentage of missing SNP data per individual of *Tyto alba alba* in museum specimens of different ages and recently collected material.

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 (Fig. 9).

GAMs relating the amount of missing data to the percentage of fragments between 35-200 bp, 200-400 bp, 400-700 bp and 700-10,380 bp 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 with the levels of missing data ($F_{1,2.3} = 32.99$, $p < 0.001$), while those at the highest bin range showed a clear negative association ($F_{1,2.4} = 37.63$, $p < 0.001$) (Fig. 10). Based on the latter model, the predicted amount of missing data when 1%, 5%, 10%, 20%, 30% or 50% of fragments ranged between 700 bp and 10,380 bp was, respectively, 88%, 77%, 65%, 42%, 23% and 4%, clearly illustrating the association between missing data and fragmentation.

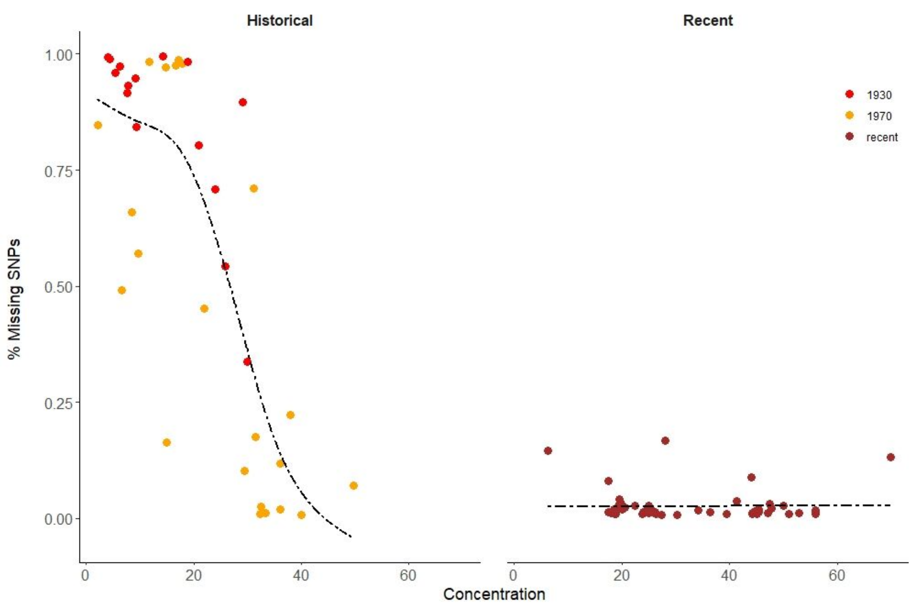

Figure 9. doi

The association between DNA concentration and percentage of missing SNPs in historical and contemporary samples of *Tyto alba alba*. The dashed line represents the predicted values according to the fitted GAM (historical samples) and linear regression (recent samples).

To date, few studies have assessed whether RRS on museum collections is feasible, and if so, how to optimise approaches. In a previous study using ddRAD target enriched sequencing, an inclusion threshold for DNA concentration of 30 ng/µl was suggested (as determined from the A260 values) (Souza et al. 2017). A similar finding emerges from our study, as the percentage of missing data was notably lower from samples with DNA extract concentrations above 30 ng/µl (Fig. 9). However, overall we note that DNA fragmentation was a better predictor for the percentage of missing SNPs and successful sequencing, compared to DNA concentration. DNA concentration was not always perfectly inversely associated with DNA fragmentation as some samples with low DNA concentration also showed low levels of DNA fragmentation, or conversely, some samples with high DNA concentration were highly fragmented. Furthermore, 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. Lastly, unlike fragmentation profiles, sample DNA concentrations are species and tissue dependent, making it difficult to set a universal threshold.

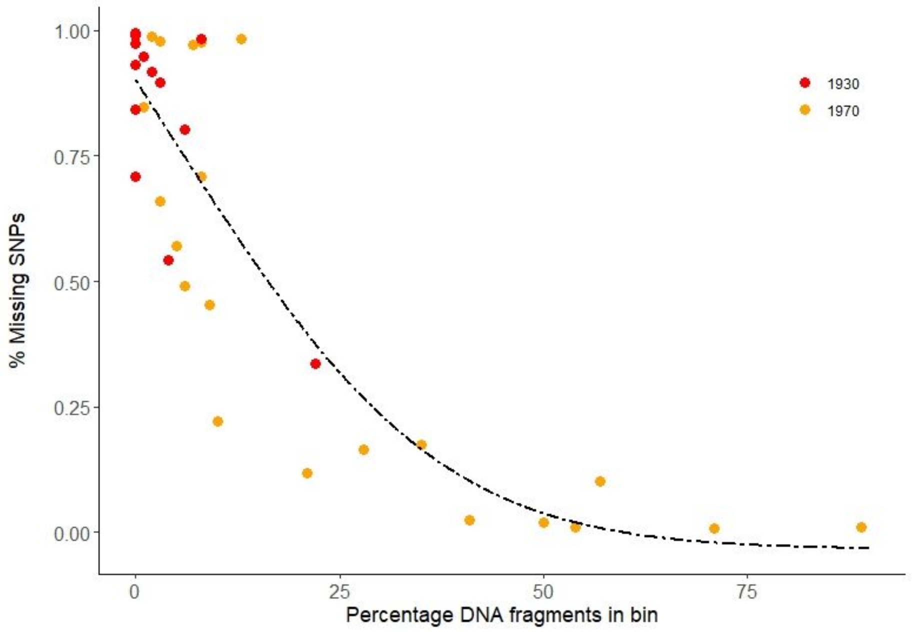

Figure 10. doi

Inverse association between the percentage of DNA fragments in the highest bin range (700-10380 bp) and percentage of missing SNPs in historical *Tyto alba alba* samples. The dashed line represents the predicted values according to the fitted GAM.

ddRAD appears unsuitable to obtain sequence data from highly-fragmented samples (in our case, the older museum samples dating from the 1930s and some more recently collected material from the 1970s). More advanced target-capture-based technologies such as HyRAD and HyRAD-X should be considered as an alternative (Suchan et al. 2016, Schmid et al. 2017), although these technologies do require additional steps and higher costs. However, obtaining population-level genomic data of museum specimens using ddRAD may still remain feasible when sufficiently large museum collections are available. Prioritising samples based on fragmentation profiles enables the targeting of effort on the most promising samples, enabling production of high-quality data in a cost-efficient manner.

Main findings and recommendations :

- ddRAD cannot be routinely applied to large museum collections to obtain population-level genomic data, especially when dealing with heavily-fragmented samples;
- However, despite the challenges of using ddRAD on degraded DNA, we were able to obtain ddRAD seq data from avian samples up to ca. 50 years old, and screening the fragment profiles of the genomic DNA gave good predictions of levels of missing data;

- Such screening is relatively easy to accomplish at minimal cost by any moderately-equipped molecular lab and substantially reduces the risk of both data loss and unnecessary library preparation and sequencing costs;
- The inclusion of data from high-quality fresh samples is important to establish a reference set to aid targeting endogenous sequence data from museum specimens.

Data availability: The raw sequencing data from this case study have been deposited at the European Nucleotide Archive (ENA) under project PRJEB59169 with accession IDs ERS14470037 - ERS14470133.

Case study 6: Single-tube DNA library preparation for ancient bones

Introduction

Massive parallel sequencing based on sequencing-by-synthesis technologies (Illumina) is an efficient method for generating DNA data from ancient material because it recovers sequences from large amounts of very short DNA fragments. In preparing samples for sequencing, single-tube DNA library protocols circumvent inter-reaction purification steps which require the transfer of DNA solutions to new tubes. They were shown to reduce DNA loss, preparation time and expenses compared to other DNA library preparation methods (Carøe et al. 2018). They also produce comparatively more-complex DNA libraries, i.e. libraries containing a higher proportion of reads mapping uniquely to the reference genome (Carøe et al. 2018). For these reasons, single-tube DNA library preparation methods represent a good option for the shotgun sequencing of ancient museum samples, especially to assess the DNA quality and quantity preserved in series of aDNA specimens. Here, we explore their application to a small series of ancient bones of Bovidae. We applied a single-tube DNA library preparation method that is based on the NEBNext Ultra Kit II DNA Library Prep Kit for Illumina (New England Biolabs) and adapted by Carøe et al. (2018) with ATDC3 adapters to avoid a uracil excision step. Since degraded DNA contains uracil residues resulting from the deamination of cytosines (Briggs et al. 2007), a uracil excision step would further fragment aDNA.

Materials and Methods

We selected seven bones of Bovidae of different ages (Table 5), from the epipalaeolithic to the late medieval (from 10,200 to 426 years old). Based on morphology, all were suspected to be aurochs (*Bos primigenius*), but some samples might be large cows (*Bos taurus*). This sampling represents a typical set of challenges for working with natural history collections where specimens are sometimes rare (wild aurochs are extinct), may have different preservation histories and may be misidentified. All manipulations took place in an aDNA lab equipped with UV lamps, under positive air pressure and following best practices recommended for working with ancient DNA (Gilbert et al. 2005, Willerslev and Cooper

2005). UV disinfection was applied before and after each experiment. Clean lab coats, masks, shoe covers and hair caps were worn for each experiment. Gloves were changed after each tube opening. Contacts with other DNA labs were banned (only sterile material was used and access to other labs was not permitted before or during the aDNA analysis). Extraction negatives (samples treated like all others, but without any bone powder inside) were included in all experiments. For tissue sampling, the outer layer of the bone was removed by scraping off its surface using a structured tooth tungsten carbide cutter attached to a hand rotary tool (8100 8v Max Rotary Tool). After 10 min of exposure to UV, 40-75 mg of bone powder was collected by drilling inside the bone fragment using the hand rotary tool at 5000 rpm, with an engraving cutter (1.6 mm). DNA was extracted from the bone powder following the protocol of Dabney et al. (2013b) and was eluted twice in 45 µl of Tris-EDTA buffer with Tween-20. For one specimen (LAST9), four separate extractions were performed. DNA extracts were then evaluated using fluorometry on a Qubit for total double stranded DNA quantification and a Bioanalyzer for fragment-size profiling.

Table 5.

Bovid bone sample information, isolated DNA concentrations and proportions of reads mapped to target and possible contaminant full genomes. ID: Tissue sample identification; conc: DNA concentration in the DNA extract measured using Qubit; mapped: percentages of the deduplicated paired-end reads mapping to the reference genomes of *Bos taurus*, *Homo sapiens* and *Mus musculus* (separated by "/"); short: percentages of mapped reads smaller than 100 bp; long: percentages of mapped reads longer than 300 bp (with insert between the paired reads); Neg1 and Neg2: negative DNA extracts processed for each library.

ID	**Epoch**	**Conc**	**Mapped**	**Short**	**Long**
		ng/µl	**% N raw**	**% < 100 bp**	**% > 300 bp**
LAST1	Roman period	1.4	0.157/0.06/0.077	83.43/2.93/1.79	0.52/10.05/10.58
LAST2	Roman period	11.8	0.644/0.01/0.004	92.87/36.08/39.93	0.31/17.67/7.72
LAST3	Roman period	2.6	1.674/0.055/0.073	85.84/14.15/5.65	0.42/6.14/9.98
LAST4	Roman period	3.7	0.02/0.03/0.054	72.39/3.7/1.88	1.38/7.71/10.1
LAST5	Roman period	5.2	0.123/0.017/0.005	88.27/11.29/19.94	0.39/26.95/9.76
LAST7	epipaleolithic	7.8	0.033/0.01/0.003	98.01/13.29/21.84	0.49/24.7/10.18
LAST9	late medieval	0.5	7.844/0.179/0.169	84.91/16.83/6.91	0.33/8.08/11.46
Neg1	NA	0	0.058/1.55/0.417	59.44/3.18/0.31	3.89/29.07/13.07
Neg3	NA	0	0.278/2.828/2.382	18.01/1.53/0.92	5.64/2.47/12.43

A total of 7 to 51 ng of genomic DNA of each specimen was used as starting material for the 'Ultra' single-tube DNA library preparation method described in Carøe et al. (2018). DNA was not sheared. The protocol consists of an end repair step, a ligation of adapters P3 and P5 (final concentration of 0.05 µM each), a fill-in reaction and purification using the MinElute kit (Qiagen). qPCR was used to evaluate DNA quantities available for each specimen for the indexing PCR. Ct values of 11.5 to 15.2 were measured by the qPCR. Based on these Ct values, 10 to 13 cycles were applied to the indexing PCR in order to perform an enrichment that would minimise impacts on the complexity of the DNA libraries

(Carøe et al. 2018). The samples were multiplexed with samples from other projects and extraction negative controls in two different libraries of six samples each and sent to Novogene (UK) Company Limited to be sequenced on an Illumina NovaSeq 6000 using a paired-end mode and 150 cycles and to produce 30 Gbp of raw data per library (Suppl. material 6). A total of 12 hours split across three days was necessary for the library preparations.

In total, 225.52 million reads (33.828 Gbp) were generated for the seven specimens and the two controls (Suppl. material 6). Illumina adapters and bad quality bases were removed using AdapterRemoval and options --trimns --trimqualities --minquality 2 (Schubert et al. 2016). Trimmed reads were assembled using PEAR (Zhang et al. 2014), mapped to the nuclear reference genomes of *Bos taurus*, *Homo sapiens* and *Mus musculus* (ARS-UCD1.2, GRCh38.p13 and GRCm39, respectively) and to the mitochondrial genome of an auroch (isolate CPC98, GenBank accession number GU985279). Duplicated reads were removed using MarkDuplicates (http://broadinstitute.github.io/picard/). Finally, post-mortem damages were evaluated for data authentication using mapDamage2 (Jónsson et al. 2013) and PMDtools (Skoglund et al. 2014). Two days of analyses were sufficient to perform the bioinformatic analyses with access to a supercomputer.

Results and Discussion

DNA concentrations after DNA extraction ranged from 0 to 11.8 ng/µl (Table 5). DNA fragment sizes showed a bimodal distribution, with one first peak below 100 bp and a second above 5000 bp. Fragments smaller than 300 bp represented more than 90% of all estimated molarity (Suppl. material 6). These should represent most of the ancient fraction of the DNA extracted. For specimen LAST9, 0.5 ng/µl DNA was detected in one extract, while no DNA was detected in the three other extracts. Two days of five hours were necessary for all DNA extractions and the price per sample was estimated at ca. €35 (including taxes, but excluding manpower). The costs of the library preparation and the sequencing were estimated at €55 and €75 per sample, respectively (including taxes, but excluding manpower). Proportions of sequenced reads mapped to the bovine genome provide an indication of the proportion of endogenous DNA and varied from 0.02% (specimen from the Roman times) to 7.84% (late medieval specimen). Compared to the reads mapped to the human or to the mouse genomes, those mapped to the bovine genome represented a higher percentage of all reads in all samples but one (LAST4). They showed a narrower size distribution, concentrated below 100 bp and shorter insert sizes (Table 5). They also showed patterns of DNA degradation in mapDamage2 (Jónsson et al. 2013) and higher post-mortem damage scores (Suppl. material 6) in PMDtools (Skoglund et al. 2014). These features are indicative of ancient endogenous DNA and were not observed in the negative controls (Table 5). Between 1.8 and 96% of the bovine mitochondrial genome was recovered, depending on the samples, with some regions covered from 0 to 32 times.

Data authentication

The single-tube library preparation protocol applied here (Carøe et al. 2018) provided DNA reads for all seven specimens tested, with varying proportions of DNA that mapped to the reference genome of *Bos taurus*. The authentication of these reads is critical for downstream analysis and should test both the ancient and endogenous origin of the reads filtered for analysis. This includes checking signatures of degradation including nucleotide alterations and DNA size profiles (Hofreiter et al. 2001, Renaud et al. 2019) and comparisons with reads from negative controls and reads mapped to other genomes. Here, most reads mapped to the bovine genome were smaller than 100 bp and in the same range as those obtained by Carøe et al. (2018) from eight historic grey wolf skins of 90 to 146 years old (40-180 bp with an average of around 60 bp). Estimating insert sizes when mapping paired reads is useful to evaluate the size distribution of DNA fragments in the library. Thus, even though it is more expensive, generating paired-end reads and reads larger than the average short read lengths revealed by the Bioanalyzer (i.e. 100 cycles or more), enables the exclusion of reads obtained from longer DNA fragments, which may correspond to recent contaminations. It is also important to filter out contaminant reads that still map to the target genome. Indeed, short contaminant fragments can map to evolutionary more conserved regions of divergent genomes. Thus, removing reads that map both to the target genome and other divergent genomes is a useful precaution. Competitive mapping can address this by mapping raw sequencing data to a concatenated reference composed of the target species' genome and other possible contaminant genomes, such as the human genome. The sequences aligned only to the target part of the concatenated reference genome can be kept for downstream analyses (Feuerborn et al. 2020). Further authentication would include a completely independent analysis (from extraction to sequencing) to check the congruence of the results (Andreeva et al. 2022).

Main findings and recommendations:

- Streamlined single-tube DNA library preparation methods adapted for degraded DNA and followed by shallow shotgun sequencing are useful for estimating the percentages of ancient endogenous DNA recoverable from DNA extracts;
- Data authentication should integrate as many aspects of the endogenous DNA as possible. Comparisons with controls (extraction negatives and unrelated genomic data) are crucial to assess the risk of including contaminant DNA in the analysis and to guide steps to filter them out;
- DNA fragment-size profiles of the DNA extracts are indicative of the presence of degraded DNA, but sequencing is necessary to evaluate percentages of endogenous DNA.

Data availability: The raw sequencing data for this case study have been deposited at the European Nucleotide Archive (ENA) under project PRJEB59185 with accession IDs ERS14471070 - ERS14471078.

Concluding remarks

The continually evolving landscape of sequencing platforms and chemistries is resulting in an ever-expanding set of opportunities for unlocking the genomic resources held in natural history collections and there is a general increase in the feasibility of museum specimen sequencing. With the rapid expansion in the field of museomics, comes a pressing need for the ongoing development, sharing and adoption of best practices. Areas of particular importance include establishment of appropriate facilities, workflows and data verification steps to minimise risks of contamination and sampling guidance which supports optimal utilisation of museum specimens for genomic research. Another area of general importance is attention to ethical issues associated with the use of specimens for genomic science; many collections pre-date contemporary permit conditions or restrictions. Guidelines for ethical issues associated with sampling specimens for genomic analysis are mostly developed for human tissues and archaeofaunal remains (Prendergast and Sawchuk 2018, Pálsdóttir et al. 2019); further dialogue (e.g. Canales et al. (2022)) and policy development regarding best practice for genomic sampling of wider natural history collections are needed.

Acknowledgements

We are grateful to Sean Prosser and Evegeny Zakharov (University of Guelph), Ben Price (Natural History Museum London), Ryan Folk (Mississippi State University) and Chris Raxworthy (American Museum of Natural History) for advice; to five reviewers and the subject editor for their comments on the paper; to Suzanne Cubey at the Royal Botanic Garden Edinburgh for facilitating access to the herbarium collection for tissue sampling; to the Department of Biosciences at the University of Oslo for access to the ancient DNA laboratory; to Bea De Cupere, Quentin Goffette, Mietje Germonpré and Annelise Folie from the Royal Belgian Institute of Natural Sciences, Belgium for designing the project “LAST”, authorising tissue sampling, giving access to RBINS specimens and to their associated information; to the Kerkuilenwerkgroep Vlaanderen for providing tissue samples; to Lejia Zhang (MfN Berlin) for isolating DNA and preparing libraries and Susan Mbedi and Sarah Sparman (Berlin Center for Genomics in Biodiversity Research) for conducting quality checks and sequencing.

Funding program

This research received support from the SYNTHESYS+ Project (http://www.synthesys.info/) which was financed by European Community Research Infrastructure Action under the H2020 Integrating Activities Programme, Project number 823827. The Royal Botanic Garden Edinburgh acknowledges support from the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS). Analyses at RBGE were run on the Crop Diversity Bioinformatics Resource which is funded by BBSRC BB/S019669/1. The Joint Experimental Molecular Unit (JEMU) of the Royal Belgian

Institute of Natural Sciences (RBINS, Brussels) and the Royal Museum for Central Africa (RMCA, Tervuren) acknowledge funding support from the Belgian Science Policy (Belspo).

Grant title

This work was funded as part of the SYNTHESYS+ Project (http://www.synthesys.info)

Author contributions

GF, LE, MLH, SJ, TvR, GS, CV, MV and PMH conceived and designed the study, GF and PMH produced the initial draft of paper and all authors contributed to augmenting, refining and revising the manuscript. Case study 1 (insect mitogenomics) was led by TvR and JP; Case study 2 (WGS of insects) by LE and MV; Case study 3 (genome skimming of plants) by SJ and MM; Case study 4 (target capture of plants) by GF, MLH, CK, FP and PMH; Case study 5 (RAD seq of birds) by CV; Case study 6 (bovid bone sequencing) by GS.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Aird D, Ross M, Chen W, Danielsson M, Fennell T, Russ C, Jaffe D, Nusbaum C, Gnirke A (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology 12 (2): R18. https://doi.org/10.1186/gb-2011-12-2-r18
- Allentoft M, Collins M, Harker D, Haile J, Oskam C, Hale M, Campos P, Samaniego J, Gilbert MTP, Willerslev E, Zhang G, Scofield RP, Holdaway R, Bunce M (2012) The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proceedings of the Royal Society: Biological Sciences 279 (1748): 4724-4733. https://doi.org/10.1098/rspb.2012.1745
- Alsos IG, Lavergne S, Merkel MKF, Boleda M, Lammers Y, Alberti A, Pouchon C, Denoeud F, Pitelkova I, Pușcaș M, Roquet C, Hurdu B, Thuiller W, Zimmermann N, Hollingsworth P, Coissac E (2020) The treasure vault can be opened: large-scale genome skimming works well using herbarium and silica gel dried aterial. Plants 9 (4): 432. https://doi.org/10.3390/plants9040432
- Andreeva TV, Malyarchuk AB, Soshkina AD, Dudko NA, Plotnikova MY, Rogaev EI (2022) Methodologies for ancient DNA extraction from bones for genomic analysis: Approaches and guidelines. Russian Journal of Genetics 58 (9): 1017-1035. https://doi.org/10.1134/s1022795422090034
- Andrews S (2010) FastQC A quality control tool for high throughput sequence data (Online). Available online at: URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc

- Austin R, Sholts S, Williams L, Kistler L, Hofman C (2019) To curate the molecular past, museums need a carefully considered set of best practices. Proceedings of the National Academy of Sciences of the United States of America 116 (5): 1471-1474. https://doi.org/10.1073/pnas.1822038116
- Bakker F, Lei D, Yu J, Mohammadin S, Wei Z, van de Kerke S, Gravendeel B, Nieuwenhuis M, Staats M, Alquezar-Planas D, Holmer R (2016) Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. Biological Journal of the Linnean Society 117 (1): 33-43. https://doi.org/10.1111/bij.12642
- Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov A, Lesin V, Nikolenko S, Pham S, Prjibelski A, Pyshkin A, Sirotkin A, Vyahhi N, Tesler G, Alekseyev M, Pevzner P (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19 (5): 455-477. https://doi.org/10.1089/cmb.2012.0021
- Bebber D, Carine M, Wood JI, Wortley A, Harris D, Prance G, Davidse G, Paige J, Pennington T, Robson NB, Scotland R (2010) Herbaria are a major frontier for species discovery. Proceedings of the National Academy of Sciences of the United States of America 107 (51): 22169-22171. https://doi.org/10.1073/pnas.1011841108
- Besnard G, Bertrand JM, Delahaie B, Bourgeois YC, Lhuillier E, Thébaud C (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): 71-82. https://doi.org/10.1111/bij.12494
- Besnard G, Gaudeul M, Lavergne S, Muller S, Rouhan G, Sukhorukov A, Vanderpoorten A, Jabbour F (2018) Herbarium-based science in the twenty-first century. Botany Letters 165 (3-4): 323-327. https://doi.org/10.1080/23818107.2018.1482783
- Bieker V, Sánchez Barreiro F, Rasmussen J, Brunier M, Wales N, Martin M (2020) Metagenomic analysis of historical herbarium specimens reveals a postmortem microbial community. Molecular Ecology Resources 20 (5): 1206-1219. https://doi.org/10.1111/1755-0998.13174
- Billerman S, Walsh J (2019) Historical DNA as a tool to address key questions in avian biology and evolution: A review of methods, challenges, applications, and future directions. Molecular Ecology Resources 19 (5): 1115-1130. https://doi.org/10.1111/1755-0998.13066
- Blaimer B, Lloyd M, Guillory W, Brady S (2016) Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect Specimens. PLOS One 11 (8): e0161531. https://doi.org/10.1371/journal.pone.0161531
- Bolger A, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30 (15): 2114-2120. https://doi.org/10.1093/bioinformatics/btu170
- Bradshaw M, Carey J, Liu M, Bartholomew H, Jurick W, Hambleton S, Hendricks D, Schnittler M, Scholler M (2023) Genetic time traveling: sequencing old herbarium specimens, including the oldest herbarium specimen sequenced from kingdom Fungi, reveals the population structure of an agriculturally significant rust. New Phytologist 237 (4): 1463-1473. https://doi.org/10.1111/nph.18622
- Brewer G, Clarkson J, Maurin O, Zuntini A, Barber V, Bellot S, Biggs N, Cowan R, Davies NJ, Dodsworth S, Edwards S, Eiserhardt W, Epitawalage N, Frisby S, Grall A,

Kersey P, Pokorny L, Leitch I, Forest F, Baker W (2019) Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. Frontiers in Plant Science 10: 1102. https://doi.org/10.3389/fpls.2019.01102

- Briggs A, Stenzel U, Johnson PF, Green R, Kelso J, Prüfer K, Meyer M, Krause J, Ronan M, Lachmann M, Pääbo S (2007) Patterns of damage in genomic DNA sequences from a Neandertal. Proceedings of the National Academy of Sciences USA 37: 14616-14621. https://doi.org/10.1073/pnas.0704665104
- Brotherton P, Endicott P, Sanchez J, Beaumont M, Barnett R, Austin J, Cooper A (2007) Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. Nucleic Acids Research 35 (17): 5717-5728. https://doi.org/10.1093/nar/gkm588
- Brutlag D, Schlehuber C, Bonner J (1969) Properties of formaldehyde-treated nucleohistone. Biochemistry 8 (8): 3214-3218. https://doi.org/10.1021/bi00836a013
- Burrell A, Disotell T, Bergey C (2015) The use of museum specimens with high-throughput DNA sequencers. Journal of Human Evolution 79: 35-44. https://doi.org/10.1016/j.jhevol.2014.10.015
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden T (2009) BLAST+: architecture and applications. BMC Bioinformatics 10: 421. https://doi.org/10.1186/1471-2105-10-421
- Camacho MA, Salgado M J, Burneo S (2018) An accounting approach to calculate the financial value of a natural history collection of mammals in Ecuador. Museum Management and Curatorship 33 (3): 279-296. https://doi.org/10.1080/09647775.2018.1466191
- Canales NA, Clarke AC, Nesbitt M, Gutaker R (2022) DNA from museum collections. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Vol. 1. Advanced Books https://doi.org/10.3897/ab.e98875
- Card D, Shapiro B, Giribet G, Moritz C, Edwards S (2021) Museum genomics. Annual Review of Genetics 55: 633-659. https://doi.org/10.1146/annurev-genet-071719-020506
- Carøe C, Gopalakrishnan S, Vinner L, Mak ST, Sinding M, Samaniego J, Wales N, Sicheritz-Pontén T, Gilbert MTP (2018) Single-tube library preparation for degraded DNA. Methods in Ecology and Evolution 9 (2): 410-419. https://doi.org/10.1111/2041-210X.12871
- Carpenter M, Buenrostro J, Valdiosera C, Schroeder H, Allentoft M, Sikora M, Rasmussen M, Gravel S, Guillén S, Nekhrizov G, Leshtakov K, Dimitrova D, Theodossiev N, Pettener D, Luiselli D, Sandoval K, Moreno-Estrada A, Li Y, Wang J, Gilbert MTP, Willerslev E, Greenleaf W, Bustamante C (2013) Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. American Journal of Human Genetics 93 (5): 852-864. https://doi.org/10.1016/j.ajhg.2013.10.002
- Carter D, Walker AK (1999) Care and Conservation of Natural History Collections. Oxford: Butterworth Heinemann URL: http://www.natsca.org/care-and-conservation
- Casbon J, Osborne R, Brenner S, Lichtenstein C (2011) A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Research 39 (12): e81. https://doi.org/10.1093/nar/gkr217

- Catchen J, Amores A, Hohenlohe P, Cresko W, Postlethwait J (2011) Stacks: building and genotyping loci de novo from short-read sequences. G3, Genes, Genomes, Genetics 1 (3): 171-182. https://doi.org/10.1534/g3.111.000240
- Cavill EL, Liu S, Zhou X, Gilbert MTP (2022) To bee, or not to bee? One leg is the question. Molecular Ecology Resources 22 (5): 1868-1874. https://doi.org/10.1111/1755-0998.13578
- Clewing C, Kehlmaier C, Stelbrink B, Albrecht C, Wilke T (2022) Poor hDNA-derived NGS data may provide sufficient phylogenetic information of potentially extinct taxa. Frontiers in Ecology and Evolution 10: 907889. https://doi.org/10.3389/fevo.2022.907889
- Colella J, Tigano A, MacManes M (2020) A linked-read approach to museomics: Higher quality de novo genome assemblies from degraded tissues. Molecular Ecology Resources 20 (4): 856-870. https://doi.org/10.1111/1755-0998.13155
- Cong Q, Shen J, Zhang J, Li W, Kinch L, Calhoun J, Warren A, Grishin N (2021) Genomics reveals the origins of historical specimens. Molecular Biology and Evolution 38 (5): 2166-2176. https://doi.org/10.1093/molbev/msab013
- Consortium of European Taxonomic Facilities (2015) CETAF Code of Conduct and Best Practice. https://www.cetaf.org/wp-content/uploads/CETAF-Code-of-Conduct-and-Best-Practice-material-transfer-agreements-2015.pdf.
- Crampton-Platt A, Yu D, Zhou X, Vogler A (2016) Mitochondrial metagenomics: letting the genes out of the bottle. GigaScience 5 (1): s13742-016-0120-y. https://doi.org/10.1186/s13742-016-0120-y
- Cridland J, Ramirez S, Dean C, Sciligo A, Tsutsui N (2018) Genome sequencing of museum specimens reveals rapid changes in the genetic composition of honey bees in California. Genome Biology and Evolution 10 (2): 458-472. https://doi.org/10.1093/gbe/evy007
- Cronn R, Liston A, Parks M, Gernandt D, Shen R, Mockler T (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Research 36 (19): e122. https://doi.org/10.1093/nar/gkn502
- Dabney J, Meyer M (2012) Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. BioTechniques 52 (2): 87-94. https://doi.org/10.2144/000113809
- Dabney J, Meyer M, Pääbo S (2013a) Ancient DNA damage. Cold Spring Harbor Perspectives in Biology 5 (7): a012567. https://doi.org/10.1101/cshperspect.a012567
- Dabney J, Knapp M, Glocke I, Gansauge M, Weihmann A, Nickel B, Valdiosera C, García N, Pääbo S, Arsuaga J, Meyer M (2013b) Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proceedings of the National Academy of Sciences of the United States of America 110 (39): 15758-15763. https://doi.org/10.1073/pnas.1314445110
- Daley T, Smith A (2014) Modeling genome coverage in single-cell sequencing. Bioinformatics 30 (22): 3159-3165. https://doi.org/10.1093/bioinformatics/btu540
- Danecek P, Auton A, Abecasis G, Albers C, Banks E, DePristo M, Handsaker R, Lunter G, Marth G, Sherry S, McVean G, Durbin R, Genomes Project Analysis G (2011) The variant call format and VCFtools. Bioinformatics 27 (15): 2156-2158. https://doi.org/10.1093/bioinformatics/btr330

- Davey J, Blaxter M (2010) RADSeq: next-generation population genetics. Briefings in Functional Genomics 9 (5-6): 416-423. https://doi.org/10.1093/bfgp/elq031
- Davey J, Hohenlohe P, Etter P, Boone J, Catchen J, Blaxter M (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12 (7): 499-510. https://doi.org/10.1038/nrg3012
- de Filippo C, Meyer M, Prüfer K (2018) Quantifying and reducing spurious alignments for the analysis of ultra-short ancient DNA sequences. BMC Biology 16: 121. https://doi.org/10.1186/s12915-018-0581-9
- de Mestier A, Mulcahy D, Harris D, Korotkova N, Long S, Häffner E, Paton A, Schiller E, Leliaert F, Mackenzie-Dodds J, Fulcher T, Stahls G, von Rintelen T, Martín M, Lücking R, Williams C, Lyal C, Güntsch A, Aronsson H, Castelin M, Pielach A, Poczai P, Ruiz León Y, Sanmartin Bastida I, Thines M, Droege G (2022) Policies Handbook on Using Molecular Collections. ARPHA Preprints https://doi.org/10.3897/arphapreprints.e98432
- Dentinger BM, Gaya E, O'Brien H, Suz L, Lachlan R, Díaz-Valderrama J, Koch R, Aime MC (2016) Tales from the crypt: genome mining from fungarium specimens improves resolution of the mushroom tree of life. Biological Journal of the Linnean Society 117 (1): 11-32. https://doi.org/10.1111/bij.12553
- D'Ercole J, Prosser SWJ, Hebert PDN (2021) A SMRT approach for targeted amplicon sequencing of museum specimens (Lepidoptera)-patterns of nucleotide misincorporation. PeerJ 9 https://doi.org/10.7717/peerj.10420
- Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin 19: 11-15.
- Espeland M, Irestedt M, Johanson KA, Akerlund M, Bergh J, Källersjö M (2010) Dichlorvos exposure impedes extraction and amplification of DNA from insects in museum collections. Frontiers in Zoology 7: 2. https://doi.org/10.1186/1742-9994-7-2
- Ewart K, Johnson R, Ogden R, Joseph L, Frankham G, Lo N (2019) Museum specimens provide reliable SNP data for population genomic analysis of a widely distributed but threatened cockatoo species. Molecular Ecology Resources 19 (6): 1578-1592. https://doi.org/10.1111/1755-0998.13082
- Ferrari G, Neukamm J, Baalsrud H, Breidenstein A, Ravinet M, Phillips C, Rühli F, Bouwman A, Schuenemann V (2020) Variola virus genome sequenced from an eighteenth-century museum specimen supports the recent origin of smallpox. Philosophical Transactions of the Royal Society B 375 (1812): 20190572. https://doi.org/10.1098/rstb.2019.0572
- Feuerborn T, Palkopoulou E, van der Valk T, von Seth J, Munters A, Pečnerová P, Dehasque M, Ureña I, Ersmark E, Lagerholm VK, Krzewińska M, Rodríguez-Varela R, Götherström A, Dalén L, Díez-Del-Molino D (2020) Competitive mapping allows for the identification and exclusion of human DNA contamination in ancient faunal genomic datasets. BMC Genomics 21: 844. https://doi.org/10.1186/s12864-020-07229-y
- Folk R, Kates H, LaFrance R, Soltis D, Soltis P, Guralnick R (2021) High-throughput methods for efficiently building massive phylogenies from natural history collections. Applications in Plant Sciences 9 (2): e11410. https://doi.org/10.1002/aps3.11410
- Forrest L, Hart M, Hughes M, Wilson H, Chung K, Tseng Y, Kidner C (2019) The Limits of Hyb-Seq for herbarium specimens: impact of preservation techniques. Frontiers in Ecology and Evolution 7: 439. https://doi.org/10.3389/fevo.2019.00439
- Freedman J, Van Dorp L, Brace S (2018) Destructive sampling natural science collections: An overview for museum professionals and researchers. Journal of Natural

Science Collections 5: 21-34. URL: https://www.natsca.org/sites/default/files/publications/JoNSC-Vol5-FreedmanVanDorpBrace2018.pdf

- Fulton T (2012) Setting Up an Ancient DNA Laboratory. In: Shapiro B, Hofreiter M (Eds) Ancient DNA: Methods and Protocols. Humana Press, Totowa, NJ, 10 pp. [ISBN 9781617795169]. https://doi.org/10.1007/978-1-61779-516-9_1
- Fu Y, Wu P, Beane T, Zamore P, Weng Z (2018) Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. BMC Genomics 19: 531. https://doi.org/10.1186/s12864-018-4933-1
- Gilbert MTP, Bandelt H, Hofreiter M, Barnes I (2005) Assessing ancient DNA studies. Trends in Ecology & Evolution 20 (10): 541-544. https://doi.org/10.1016/j.tree.2005.07.005
- Gilbert MTP, Moore W, Melchior L, Worobey M (2007a) DNA extraction from dry museum beetles without conferring external morphological damage. PLoS ONE 2 (3): e272. https://doi.org/10.1371/journal.pone.0000272
- Gilbert MTP, Haselkorn T, Bunce M, Sanchez J, Lucas S, Jewell L, Van Marck E, Worobey M (2007b) The isolation of nucleic acids from fixed, paraffin-embedded tissues–Which methods are useful when? PLoS ONE 2 (6): e537. https://doi.org/10.1371/journal.pone.0000537
- Graham C, Glenn T, McArthur A, Boreham D, Kieran T, Lance S, Manzon R, Martino J, Pierson T, Rogers S, Wilson J, Somers C (2015) Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). Molecular Ecology Resources 15 (6): 1304-1315. https://doi.org/10.1111/1755-0998.12404
- Green R, Briggs A, Krause J, Prüfer K, Burbano H, Siebauer M, Lachmann M, Pääbo S (2009) The Neandertal genome and ancient DNA authenticity. The EMBO Journal 28 (17): 2494-2502. https://doi.org/10.1038/emboj.2009.222
- Gutaker R, Weiß C, Ellis D, Anglin N, Knapp S, Luis Fernández-Alonso J, Prat S, Burbano H (2019) The origins and adaptation of European potatoes reconstructed from historical genomes. Nature Ecology and Evolution 3 (7): 1093-1101. https://doi.org/10.1038/s41559-019-0921-3
- Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - a baiting and iterative mapping approach. Nucleic Acids Research 41 (13): e129. https://doi.org/10.1093/nar/gkt371
- Hahn E, Alexander M, Grealy A, Stiller J, Gardiner D, Holleley C (2021) Unlocking inaccessible historical genomes preserved in formalin. Molecular Ecology Resources 22 (6): 2130-2147. https://doi.org/10.1111/1755-0998.13505
- Hart M, Forrest L, Nicholls J, Kidner C (2016) Retrieval of hundreds of nuclear loci from herbarium specimens. Taxon 65 (5): 1081-1092. https://doi.org/10.12705/655.9
- Hebert PN, Dewaard J, Zakharov E, Prosser SJ, Sones J, McKeown JA, Mantle B, La Salle J (2013) A DNA 'barcode blitz': rapid digitization and sequencing of a natural history collection. PLOS One 8 (7): e68535. https://doi.org/10.1371/journal.pone.0068535
- Heintzman P, Elias S, Moore K, Paszkiewicz K, Barnes I (2014) Characterizing DNA preservation in degraded specimens of *Amara alpina* (Carabidae: Coleoptera). Molecular Ecology Resources 14 (3): 606-615. https://doi.org/10.1111/1755-0998.12205

- Hobern D (2021) BIOSCAN: DNA barcoding to accelerate taxonomy and biogeography for conservation and sustainability. Genome 64 (3): 161-164. https://doi.org/10.1139/gen-2020-0009
- Hodge WH (1947) The use of alcohol in plant collecting. Rhodora 49 (584): 207-210. URL: http://www.jstor.org/stable/23303840
- Hofreiter M, Jaenicke V, Serre D, von Haeseler A, Pääbo S (2001) DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic Acids Research 29 (23): 4793-4799. https://doi.org/10.1093/nar/29.23.4793
- Holmes M, Hammond T, Wogan GU, Walsh R, LaBarbera K, Wommack E, Martins F, Crawford J, Mack K, Bloch L, Nachman M (2016) Natural history collections as windows on evolutionary processes. Molecular Ecology 25 (4): 864-881. https://doi.org/10.1111/mec.13529
- Irestedt M, Thörn F, Müller I, Jønsson K, Ericson PP, Blom MK (2022) A guide to avian museomics: Insights gained from resequencing hundreds of avian study skins. Molecular Ecology Resources 22 (7): 2672-2684. https://doi.org/10.1111/1755-0998.13660
- Jensen E, Díez-Del-Molino D, Gilbert MTP, Bertola L, Borges F, Cubric-Curik V, de Navascués M, Frandsen P, Heuertz M, Hvilsom C, Jiménez-Mena B, Miettinen A, Moest M, Pečnerová P, Barnes I, Vernesi C (2022) Ancient and historical DNA in conservation policy. Trends in Ecology & Evolution 37 (5): 420-429. https://doi.org/10.1016/j.tree.2021.12.010
- Jin J, Yu W, Yang J, Song Y, dePamphilis C, Yi T, Li D (2020) GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biology 21: 241. https://doi.org/10.1186/s13059-020-02154-5
- Jin M, Zwick A, Ślipiński A, Keyzer R, Pang H (2020) Museomics reveals extensive cryptic diversity of Australian prionine longhorn beetles with implications for their classification and conservation. Systematic Entomology 45 (4): 745-770. https://doi.org/10.1111/syen.12424
- Johnson K, Owens IP, The Global Collections Group (2023) A global approach for natural history museum collections. Science 379 (6638): 1192-1194. https://doi.org/10.1126/science.adf6434
- Johnson M, Pokorny L, Dodsworth S, Botigué L, Cowan R, Devault A, Eiserhardt W, Epitawalage N, Forest F, Kim J, Leebens-Mack J, Leitch I, Maurin O, Soltis D, Soltis P, Wong GK, Baker W, Wickett N (2019) A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. Systematic Biology 68 (4): 594-606. https://doi.org/10.1093/sysbio/syy086
- Jónsson H, Ginolhac A, Schubert M, Johnson PF, Orlando L (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 29 (13): 1682-1684. https://doi.org/10.1093/bioinformatics/btt193
- Kapp J, Green R, Shapiro B (2021) A fast and efficient single-stranded genomic library preparation method optimized for Ancient DNA. Journal of Heredity 112 (3): 241-249. https://doi.org/10.1093/jhered/esab012
- Kates H, Doby J, Siniscalchi C, LaFrance R, Soltis D, Soltis P, Guralnick R, Folk R (2021) The effects of herbarium specimen characteristics on short-read NGS sequencing success in nearly 8000 specimens: old, degraded samples have lower DNA

yields but consistent sequencing success. Frontiers in Plant Science 12: 669064. https://doi.org/10.3389/fpls.2021.669064

- Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30 (14): 3059-3066. https://doi.org/10.1093/nar/gkf436
- Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28 (12): 1647-1649. https://doi.org/10.1093/bioinformatics/bts199
- Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Research 40 (1): e3. https://doi.org/10.1093/nar/gkr771
- Kistler L, Ware R, Smith O, Collins M, Allaby R (2017) A new model for ancient DNA decay based on paleogenomic meta-analysis. Nucleic Acids Research 45 (11): 6310-6320. https://doi.org/10.1093/nar/gkx361
- Kjær K, Winther Pedersen M, De Sanctis B, De Cahsan B, Korneliussen T, Michelsen C, Sand K, Jelavić S, Ruter A, Schmidt AA, Kjeldsen K, Tesakov A, Snowball I, Gosse J, Alsos I, Wang Y, Dockter C, Rasmussen M, Jørgensen M, Skadhauge B, Prohaska A, Kristensen JÅ, Bjerager M, Allentoft M, Coissac E, PhyloNorway C, Rouillard A, Simakova A, Fernandez-Guerra A, Bowler C, Macias-Fauria M, Vinner L, Welch J, Hidy A, Sikora M, Collins M, Durbin R, Larsen N, Willerslev E (2022) A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. Nature 612 (7939): 283-291. https://doi.org/10.1038/s41586-022-05453-y
- Knapp M, Clarke A, Horsburgh KA, Matisoo-Smith E (2012) Setting the stage - building and working in an ancient DNA laboratory. Annals of Anatomy 194 (1): 3-6. https://doi.org/10.1016/j.aanat.2011.03.008
- Knyshov A, Gordon EL, Weirauch C (2019) Cost-efficient high throughput capture of museum arthropod specimen DNA using PCR-generated baits. Methods in Ecology and Evolution 10 (6): 841-852. https://doi.org/10.1111/2041-210x.13169
- Korlević P, McAlister E, Mayho M, Makunin A, Flicek P, Lawniczak MN (2021) A minimally morphologically destructive approach for DNA retrieval and whole genome shotgun sequencing of pinned historic Dipteran vector species. Genome Biology and Evolution 13 (10): evab226. https://doi.org/10.1093/gbe/evab226
- Krause J, Briggs A, Kircher M, Maricic T, Zwyns N, Derevianko A, Pääbo S (2010) A complete mtDNA genome of an early modern human from Kostenki, Russia. Current Biology 20 (3): 231-236. https://doi.org/10.1016/j.cub.2009.11.068
- Lachance J, Tishkoff S (2013) SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. Bioessays 35 (9): 780-786. https://doi.org/10.1002/bies.201300014
- Lalueza-Fox C (2022) Museomics. Current Biology 32 (21): R1214-R1215. https://doi.org/10.1016/j.cub.2022.09.019
- Langmead B, Salzberg S (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9 (4): 357-359. https://doi.org/10.1038/nmeth.1923
- Lang PM, Weiß C, Kersten S, Latorre S, Nagel S, Nickel B, Meyer M, Burbano H (2020) Hybridization ddRAD-sequencing for population genomics of nonmodel plants using

highly degraded historical specimen DNA. Molecular Ecology Resources 20 (5): 1228-1247. https://doi.org/10.1111/1755-0998.13168

- Latorre S, Lang PM, Burbano H, Gutaker R (2020) Isolation, library preparation, and bioinformatic analysis of historical and ancient plant DNA. Current Protocols in Plant Biology 5: e20121. https://doi.org/10.1002/cppb.20121
- Levesque-Beaudin V, Miller M, Dikow T, Miller S, Prosser SJ, Zakharov E, McKeown JA, Sones J, Redmond N, Coddington J, Santos B, Bird J, deWaard J (2023) A workflow for expanding DNA barcode reference libraries through 'museum harvesting' of natural history collections. Biodiversity Data Journal 11: e100677. https://doi.org/10.3897/BDJ.11.e100677
- Lewin H, Richards S, Lieberman Aiden E, Allende M, Archibald J, Bálint M, Barker K, Baumgartner B, Belov K, Bertorelle G, Blaxter M, Cai J, Caperello N, Carlson K, Castilla-Rubio JC, Chaw S, Chen L, Childers A, Coddington J, Conde D, Corominas M, Crandall K, Crawford A, DiPalma F, Durbin R, Ebenezer T, Edwards S, Fedrigo O, Flicek P, Formenti G, Gibbs R, Gilbert MTP, Goldstein M, Graves JM, Greely H, Grigoriev I, Hackett K, Hall N, Haussler D, Helgen K, Hogg C, Isobe S, Jakobsen KS, Janke A, Jarvis E, Johnson W, Jones SM, Karlsson E, Kersey P, Kim J, Kress WJ, Kuraku S, Lawniczak MN, Leebens-Mack J, Li X, Lindblad-Toh K, Liu X, Lopez J, Marques-Bonet T, Mazard S, Mazet JK, Mazzoni C, Myers E, O'Neill R, Paez S, Park H, Robinson G, Roquet C, Ryder O, Sabir JM, Shaffer HB, Shank T, Sherkow J, Soltis P, Tang B, Tedersoo L, Uliano-Silva M, Wang K, Wei X, Wetzer R, Wilson J, Xu X, Yang H, Yoder A, Zhang G (2022) The Earth BioGenome Project 2020: Starting the clock. Proceedings of the National Academy of Sciences of the United States of America 119 (4): e2115635118. https://doi.org/10.1073/pnas.2115635118
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25 (14): 1754-1760. https://doi.org/10.1093/bioinformatics/btp324
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25 (16): 2078-2079. https://doi.org/10.1093/bioinformatics/btp352
- Li H, Luo Y, Gan L, Ma P, Gao L, Yang J, Cai J, Gitzendanner M, Fritsch P, Zhang T, Jin J, Zeng C, Wang H, Yu W, Zhang R, van der Bank M, Olmstead R, Hollingsworth P, Chase M, Soltis D, Soltis P, Yi T, Li D (2021) Plastid phylogenomic insights into relationships of all flowering plant families. BMC Biology 19: 232. https://doi.org/10.1186/s12915-021-01166-2
- Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362 (6422): 709-715. https://doi.org/10.1038/362709a0
- Lindgreen S (2012) AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Research Notes 5: 337. https://doi.org/10.1186/1756-0500-5-337
- Llamas B, Valverde G, Fehren-Schmitz L, Weyrich L, Cooper A, Haak W (2017) From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. STAR: Science & Technology of Archaeological Research 3 (1): 1-14. https://doi.org/10.1080/20548923.2016.1258824
- Malakasi P, Bellot S, Dee R, Grace O (2019) Museomics clarifies the classification of *Aloidendron* (Asphodelaceae), the iconic African tree Aloes. Frontiers in Plant Science 10: 1227. https://doi.org/10.3389/fpls.2019.01227

- Marx V (2017) How to deduplicate PCR. Nature Methods 14 (5): 473-476. https://doi.org/10.1038/nmeth.4268
- Mascarello M, Amalfi M, Asselman P, Smets E, Hardy O, Beeckman H, Janssens S (2021) Genome skimming reveals novel plastid markers for the molecular identification of illegally logged African timber species. PLOS One 16 (6): e0251655. https://doi.org/10.1371/journal.pone.0251655
- Matasci N, Hung L, Yan Z, Carpenter E, Wickett N, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M, Burleigh JG, Gitzendanner M, Wafula E, Der J, dePamphilis C, Roure B, Philippe H, Ruhfel B, Miles N, Graham S, Mathews S, Surek B, Melkonian M, Soltis D, Soltis P, Rothfels C, Pokorny L, Shaw J, DeGironimo L, Stevenson D, Villarreal JC, Chen T, Kutchan T, Rolf M, Baucom R, Deyholos M, Samudrala R, Tian Z, Wu X, Sun X, Zhang Y, Wang J, Leebens-Mack J, Wong GK (2014) Data access for the 1,000 Plants (1KP) project. GigaScience 3 (1): 2047-217X-3-17. https://doi.org/10.1186/2047-217X-3-17
- Mayer C, Dietz L, Call E, Kukowka S, Martin S, Espeland M (2021) Adding leaves to the Lepidoptera tree: capturing hundreds of nuclear genes from old museum specimens. Systematic Entomology 46 (3): 649-671. https://doi.org/10.1111/syen.12481
- McCormack J, Tsai WE, Faircloth B (2016) Sequence capture of ultraconserved elements from bird museum specimens. Molecular Ecology Resources 16 (5): 1189-1203. https://doi.org/10.1111/1755-0998.12466
- McDonough M, Parker L, Rotzel McInerney N, Campana M, Maldonado J (2018) Performance of commonly requested destructive museum samples for mammalian genomic studies. Journal of Mammalogy 99 (4): 789-802. https://doi.org/10.1093/jmammal/gyy080
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo M (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20 (9): 1297-1303. https://doi.org/10.1101/gr.107524.110
- Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols 2010 (6): pdb-prot5448. https://doi.org/10.1101/pdb.prot5448
- Mikheyev A, Zwick A, Magrath ML, Grau M, Qiu L, Su YN, Yeates D (2017) Museum genomics confirms that the Lord Howe Island stick insect survived extinction. Current Biology 27 (20): 3157-3161. https://doi.org/10.1016/j.cub.2017.08.058
- Miller S, Barrow L, Ehlman S, Goodheart J, Greiman S, Lutz H, Misiewicz T, Smith S, Tan M, Thawley C, Cook J, Light J (2020) Building natural history collections for the twenty-first century and beyond. BioScience 70 (8): 674-687. https://doi.org/10.1093/biosci/biaa069
- Mullin V, Stephen W, Arce A, Nash W, Raine C, Notton D, Whiffin A, Blagderov V, Gharbi K, Hogan J, Hunter T, Irish N, Jackson S, Judd S, Watkins C, Haerty W, Ollerton J, Brace S, Gill R, Barnes I (2023) First large-scale quantification study of DNA preservation in insects from natural history collections using genome-wide sequencing. Methods in Ecology and Evolution 14 (2): 360-371. https://doi.org/10.1111/2041-210x.13945
- Nadeau N, Ruiz M, Salazar P, Counterman B, Medina JA, Ortiz-Zuazaga H, Morrison A, McMillan WO, Jiggins C, Papa R (2014) Population genomics of parallel hybrid zones in

the mimetic butterflies, *H. melpomene* and *H. erato*. Genome Research 24 (8): 1316-1333. https://doi.org/10.1101/gr.169292.113

- Nakahama N (2021) Museum specimens: An overlooked and valuable material for conservation genetics. Ecological Research 36 (1): 13-23. https://doi.org/10.1111/1440-1703.12181
- Nellemann C (2012) Green Carbon, Black Trade: Illegal Logging, Tax Fraud and Laundering in the Worlds Tropical Forests. A Rapid Response Assessment. United Nations Environment Programme, GRIDArendal [ISBN 9788277011028]
- Nevill P, Zhong X, Tonti-Filippini J, Byrne M, Hislop M, Thiele K, van Leeuwen S, Boykin L, Small I (2020) Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. Plant Methods 16: 1. https://doi.org/10.1186/s13007-019-0534-5
- Nishii K, Möller M, Foster R, Forrest L, Kelso N, Barber S, Howard C, Hart M (2023) A high quality, high molecular weight DNA extraction method for PacBio HiFi genome sequencing of recalcitrant plants. Plant Methods 19 (1). https://doi.org/10.1186/s13007-023-01009-x
- Orlando L, Allaby R, Skoglund P, Der Sarkissian C, Stockhammer P, Ávila-Arcos M, Fu Q, Krause J, Willerslev E, Stone A, Warinner C (2021) Ancient DNA analysis. Nature Reviews Methods Primers 1 (1): 1-26. https://doi.org/10.1038/s43586-020-00011-0
- Pálsdóttir AH, Bläuer A, Rannamäe E, Boessenkool S, Hallsson JH (2019) Not a limitless resource: ethics and guidelines for destructive sampling of archaeofaunal remains. Royal Society Open Science 6 (10): 191059. https://doi.org/10.1098/rsos.191059
- Patzold F, Zilli A, Hundsdoerfer A (2020) Advantages of an easy-to-use DNA extraction method for minimal-destructive analysis of collection specimens. PLOS One 15 (7): e0235222. https://doi.org/10.1371/journal.pone.0235222
- Peterson B, Weber J, Kay E, Fisher H, Hoekstra H (2012) Double digest RADseq: an inexpensive method for De Novo SNP discovery and genotyping in model and non-model species. PLOS One 7 (5): e37135. https://doi.org/10.1371/journal.pone.0037135
- Peyrégne S, Prüfer K (2020) Present-day DNA contamination in ancient DNA datasets. Bioessays 42 (9): 2000081. https://doi.org/10.1002/bies.202000081
- Pinheiro J, Bates D, Team RC (2022) nlme: Linear and nonlinear mixed effects Models. R package version 3.1-160. https://svn.r-project.org/R-packages/trunk/nlme/.
- Prendergast M, Sawchuk E (2018) Boots on the ground in Africa's ancient DNA ‘revolution’: archaeological perspectives on ethics and best practices. Antiquity 92 (363): 803-815. https://doi.org/10.15184/aqy.2018.70
- Prosser SJ, deWaard J, Miller S, Hebert PN (2016) DNA barcodes from century-old type specimens using next-generation sequencing. Molecular Ecology Resources 16 (2): 487-497. https://doi.org/10.1111/1755-0998.12474
- Puritz J, Matz M, Toonen R, Weber J, Bolnick D, Bird C (2014) Demystifying the RAD fad. Molecular Ecology 23 (24): 5937-5942. https://doi.org/10.1111/mec.12965
- Ratnasingham S, Hebert PN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes 7 (3): 355-364. https://doi.org/10.1111/j.1471-8286.2007.01678.x
- Ratnasingham S, Hebert PN (2013) A DNA-based registry for all animal species: the barcode index number (BIN) system. PLOS One 8 (7): e66213. https://doi.org/10.1371/journal.pone.0066213

- Raxworthy C, Smith BT (2021) Mining museums for historical DNA: advances and challenges in museomics. Trends in Ecology & Evolution 36 (11): 1049-1060. https://doi.org/10.1016/j.tree.2021.07.009
- Rayo E, Ferrari G, Neukamm J, Akgül G, Breidenstein A, Cooke M, Phillips C, Bouwman A, Rühli F, Schuenemann V (2022) Non-destructive extraction of DNA from preserved tissues in medical collections. BioTechniques 72 (2): 60-64. https://doi.org/10.2144/btn-2021-0014
- Renaud G, Schubert M, Sawyer S, Orlando L (2019) Authentication and assessment of contamination in ancient DNA. Methods in Molecular Biology 1963: 163-194. https://doi.org/10.1007/978-1-4939-9176-1_17
- Ristaino J (2020) The importance of mycological and plant herbaria in tracking plant killers. Frontiers in Ecology and Evolution 7: 521. https://doi.org/10.3389/fevo.2019.00521
- Rochette N, Rivera-Colón A, Walsh J, Sanger T, Campbell-Staton S, Catchen J (2023) On the causes, consequences, and avoidance of PCR duplicates: towards a theory of library complexity. Molecular Ecology Resources 23: 1299-1318. https://doi.org/10.1111/1755-0998.13800
- Rohland N, Siedel H, Hofreiter M (2004) Nondestructive DNA extraction method for mitochondrial DNA analyses of museum specimens. BioTechniques 36 (5): 814-821. https://doi.org/10.2144/04365ST05
- Roycroft E, Moritz C, Rowe K, Moussalli A, Eldridge MB, Portela Miguez R, Piggott M, Potter S (2022) Sequence capture from historical museum specimens: maximizing value for population and phylogenomic studies. Frontiers in Ecology and Evolution 10: 931644. https://doi.org/10.3389/fevo.2022.931644
- Ruiz-Gartzia I, Lizano E, Marques-Bonet T, Kelley J (2022) Recovering the genomes hidden in museum wet collections. Molecular Ecology Resources 22 (6): 2127-2129. https://doi.org/10.1111/1755-0998.13631
- Sánchez Barreiro F, Vieira F, Martin M, Haile J, Gilbert MTP, Wales N (2017) Characterizing restriction enzyme-associated loci in historic ragweed (*Ambrosia artemisiifolia*) voucher specimens using custom-designed RNA probes. Molecular Ecology Resources 17 (2): 209-220. https://doi.org/10.1111/1755-0998.12610
- Santos B, Miller M, Miklasevskaja M, McKeown JA, Redmond N, Coddington J, Bird J, Miller S, Smith A, Brady S, Buffington M, Chamorro ML, Dikow T, Gates M, Goldstein P, Konstantinov A, Kula R, Silverson N, Solis MA, deWaard S, Naik S, Nikolova N, Pentinsaari M, Prosser SJ, Sones J, Zakharov E, deWaard J (2023) Enhancing DNA barcode reference libraries by harvesting terrestrial arthropods at the National Museum of Natural History. Biodiversity Data Journal 11: e100904. https://doi.org/10.3897/BDJ.11.e100904
- Särkinen T, Staats M, Richardson J, Cowan R, Bakker F (2012) How to open the treasure chest? Optimising DNA extraction from herbarium specimens. PLOS One 7 (8): e43808. https://doi.org/10.1371/journal.pone.0043808
- Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S (2012) Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. PLOS One 7 (3): e34131. https://doi.org/10.1371/journal.pone.0034131
- Schmid S, Genevest R, Gobet E, Suchan T, Sperisen C, Tinner W, Alvarez N (2017) Hy RAD -X, a versatile method combining exome capture and RAD sequencing to extract

genomic information from ancient DNA. Methods in Ecology and Evolution 8 (10): 1374-1388. https://doi.org/10.1111/2041-210x.12785

- Schubert M, Ermini L, Der Sarkissian C, Jónsson H, Ginolhac A, Schaefer R, Martin M, Fernández R, Kircher M, McCue M, Willerslev E, Orlando L (2014) Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nature Protocols 9 (5): 1056-1082. https://doi.org/10.1038/nprot.2014.063
- Schubert M, Lindgreen S, Orlando L (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Research Notes 9: 88. https://doi.org/10.1186/s13104-016-1900-2
- Seguin-Orlando A, Hoover C, Vasiliev S, Ovodov N, Shapiro B, Cooper A, Rubin E, Willerslev E, Orlando L (2015) Amplification of TruSeq ancient DNA libraries with AccuPrime Pfx: consequences on nucleotide misincorporation and methylation patterns. STAR: Science & Technology of Archaeological Research 1 (1): 1-9. https://doi.org/10.1179/2054892315Y.0000000005
- Shepherd L (2017) A non-destructive DNA sampling technique for herbarium specimens. PLOS One 12 (8): e0183555. https://doi.org/10.1371/journal.pone.0183555
- Shi L, Chen H, Jiang M, Wang L, Wu X, Huang L, Liu C (2019) CPGAVAS2, an integrated plastome sequence annotator and analyzer. Nucleic Acids Research 47 (W1): W65-W73. https://doi.org/10.1093/nar/gkz345
- Short A, Dikow T, Moreau C (2018) Entomological collections in the age of big data. Annual Review of Entomology 63: 513-530. https://doi.org/10.1146/annurev-ento-031616-035536
- Skoglund P, Northoff B, Shunkov M, Derevianko A, Pääbo S, Krause J, Jakobsson M (2014) Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proceedings of the National Academy of Sciences of the United States of America 111 (6): 2229-2234. https://doi.org/10.1073/pnas.1318934111
- Souza C, Murphy N, Villacorta-Rath C, Woodings L, Ilyushkina I, Hernandez C, Green B, Bell J, Strugnell J (2017) Efficiency of ddRAD target enriched sequencing across spiny rock lobster species (Palinuridae: Jasus). Scientific Reports 7: 6781. https://doi.org/10.1038/s41598-017-06582-5
- Souza HAV, Muller LAC, Brandão RL, Lovato MB (2012) Isolation of high quality and polysaccharide-free DNA from leaves of *Dimorphandra mollis* (Leguminosae), a tree from the Brazilian Cerrado. Genetic and Molecular Research 11 (1): 756-764. https://doi.org/10.4238/2012.March.22.6
- Soza V, Lindsley D, Waalkes A, Ramage E, Patwardhan R, Burton J, Adey A, Kumar A, Qiu R, Shendure J, Hall B (2019) The *Rhododendron* genome and chromosomal organization provide insight into shared whole-genome duplications across the heath family (Ericaceae). Genome Biology and Evolution 11 (12): 3353-3371. https://doi.org/10.1093/gbe/evz245
- Speer K, Hawkins MR, Flores M, McGowen M, Fleischer R, Maldonado J, Campana M, Muletz-Wolz C (2022) A comparative study of RNA yields from museum specimens, including an optimized protocol for extracting RNA from formalin-fixed specimens. Frontiers in Ecology and Evolution 10: 953131. https://doi.org/10.3389/fevo.2022.953131

- Staats M, Cuenca A, Richardson J, Vrielink-van Ginkel R, Petersen G, Seberg O, Bakker F (2011) DNA damage in plant herbarium tissue. PLOS One 6 (12): e28448. https://doi.org/10.1371/journal.pone.0028448
- Straube N, Lyra M, Paijmans JA, Preick M, Basler N, Penner J, Rödel M, Westbury M, Haddad CB, Barlow A, Hofreiter M (2021) Successful application of ancient DNA extraction and library construction protocols to museum wet collection specimens. Molecular Ecology Resources 21 (7): 2299-2315. https://doi.org/10.1111/1755-0998.13433
- Strijk J, Binh HT, Van Ngoc N, Pereira J, Slik JWF, Sukri R, Suyama Y, Tagane S, Wieringa J, Yahara T, Hinsinger D (2020) Museomics for reconstructing historical floristic exchanges: Divergence of stone oaks across Wallacea. PLOS One 15 (5): e0232936. https://doi.org/10.1371/journal.pone.0232936
- Suchan T, Pitteloud C, Gerasimova N, Kostikova A, Schmid S, Arrigo N, Pajkovic M, Ronikier M, Alvarez N (2016) Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. PLOS One 11 (3): e0151651. https://doi.org/10.1371/journal.pone.0151651
- Sugita N, Ebihara A, Hosoya T, Jinbo U, Kaneko S, Kurosawa T, Nakae M, Yukawa T (2020) Non-destructive DNA extraction from herbarium specimens: a method particularly suitable for plants with small and fragile leaves. Journal of Plant Research 133 (1): 133-141. https://doi.org/10.1007/s10265-019-01152-4
- Tarbet Hust ES, Snow M (2021) The effects of soft tissue removal methods on porcine skeletal remains. New Florida Journal of Anthropology 1 (2): 30-46. https://doi.org/10.32473/nfja.v1i2.124117
- The Darwin Tree of Life Project Consortium (2022) Sequence locally, think globally: The Darwin Tree of Life Project. Proceedings of the National Academy of Sciences of the United States of America 119 (4): e2115642118. https://doi.org/10.1073/pnas.2115642118
- Timmermans MTN, Viberg C, Martin G, Hopkins K, Vogler A (2016) Rapid assembly of taxonomically validated mitochondrial genomes from historical insect collections. The Biological Journal of the Linnean Society 117 (1): 83-95. https://doi.org/10.1111/bij.12552
- Töpfer T, Gamauf A, Haring E (2011) Utility of arsenic-treated bird skins for DNA extraction. BMC Research Notes 4: 197. https://doi.org/10.1186/1756-0500-4-197
- Twort V, Minet J, Wheat C, Wahlberg N (2021) Museomics of a rare taxon: placing Whalleyanidae in the Lepidoptera Tree of Life. Systematic Entomology 46 (4): 926-937. https://doi.org/10.1111/syen.12503
- Van Belleghem S, Vangestel C, De Wolf K, De Corte Z, Möst M, Rastas P, De Meester L, Hendrickx F (2018) Evolution at two time frames: Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. PLOS Genetics 14 (11): e1007796. https://doi.org/10.1371/journal.pgen.1007796
- van der Valk T, Vezzi F, Ormestad M, Dalén L, Guschanski K (2020) Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Molecular Ecology Resources 20 (5): 1171-1181. https://doi.org/10.1111/1755-0998.13009
- van der Valk T, Pečnerová P, Díez-Del-Molino D, Bergström A, Oppenheimer J, Hartmann S, Xenikoudakis G, Thomas J, Dehasque M, Sağlıcan E, Fidan FR, Barnes I, Liu S, Somel M, Heintzman P, Nikolskiy P, Shapiro B, Skoglund P, Hofreiter M, Lister A, Götherström A, Dalén L (2021) Million-year-old DNA sheds light on the genomic history

of mammoths. Nature 591 (7849): 265-269. https://doi.org/10.1038/s41586-021-03224-9

- Vuissoz A, Worobey M, Odegaard N, Bunce M, Machado C, Lynnerup N, Peacock E, Gilbert MTP (2007) The survival of PCR-amplifiable DNA in cow leather. Journal of Archaeological Science 34 (5): 823-829. https://doi.org/10.1016/j.jas.2006.09.002
- Wandeler P, Hoeck PA, Keller L (2007) Back to the future: museum specimens in population genetics. Trends in Ecology & Evolution 22 (12): 634-642. https://doi.org/10.1016/j.tree.2007.08.017
- Webster M (2017) The Extended Specimen: Emerging Frontiers in Collections-Based Ornithological Research. CRC Press URL: https://play.google.com/store/books/details?id=sjsPEAAAQBAJ [ISBN 9781498729161]
- Weiß C, Schuenemann V, Devos J, Shirsekar G, Reiter E, Gould B, Stinchcombe J, Krause J, Burbano H (2016) Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. Royal Society Open Science 3 (6): 160239. https://doi.org/10.1098/rsos.160239
- Wheeler QD, Knapp S, Stevenson DW, Stevenson J, Blum SD, Boom BM, Borisy GG, Buizer JL, De Carvalho MR, Cibrian A, Donoghue MJ, Doyle V, Gerson EM, Graham CH, Graves P, Graves SJ, Guralnick RP, Hamilton AL, Hanken J, Law W, Lipscomb DL, Lovejoy TE, Miller H, Miller JS, Naeem S, Novacek MJ, Page LM, Platnick NI, Porter-Morgan H, Raven PH, Solis MA, Valdecasas AG, Van Der Leeuw S, Vasco A, Vermeulen N, Vogel J, Walls RL, Wilson EO, Woolley JB (2012) Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. Systematics and Biodiversity 10 (1): 1-20. https://doi.org/10.1080/14772000.2012.665095
- Willerslev E, Cooper A (2005) Ancient DNA. Proceedings of the Royal Society B 272 (1558): 3-16. https://doi.org/10.1098/rspb.2004.2813
- Willerslev E, Cappellini E, Boomsma W, Nielsen R, Hebsgaard M, Brand T, Hofreiter M, Bunce M, Poinar H, Dahl-Jensen D, Johnsen S, Steffensen JP, Bennike O, Schwenninger J, Nathan R, Armitage S, de Hoog C, Alfimov V, Christl M, Beer J, Muscheler R, Barker J, Sharp M, Penkman KH, Haile J, Taberlet P, Gilbert MTP, Casoli A, Campani E, Collins M (2007) Ancient biomolecules from deep ice cores reveal a forested southern Greenland. Science 317 (5834): 111-114. https://doi.org/10.1126/science.1141758
- Wood S (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society Series B 73 (1): 3-36. https://doi.org/10.1111/j.1467-9868.2010.00749.x
- Yeates D, Zwick A, Mikheyev A (2016) Museums are biobanks: unlocking the genetic potential of the three billion specimens in the world's biological collections. Current Opinion in Insect Science 18: 83-88. https://doi.org/10.1016/j.cois.2016.09.009
- Yeo D, Srivathsan A, Meier R (2020) Longer is not always better: optimizing barcode length for large-scale species discovery and identification. Systematic Biology 69 (5): 999-1015. https://doi.org/10.1093/sysbio/syaa014
- Yoshida K, Schuenemann V, Cano L, Pais M, Mishra B, Sharma R, Lanz C, Martin F, Kamoun S, Krause J, Thines M, Weigel D, Burbano H (2013) The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. Elife 2: e00731. https://doi.org/10.7554/eLife.00731

- Zeng C, Hollingsworth P, Yang J, He Z, Zhang Z, Li D, Yang J (2018) Genome skimming herbarium specimens for DNA barcoding and phylogenomics. Plant Methods 14: 43. https://doi.org/10.1186/s13007-018-0300-0
- Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30 (5): 614-620. https://doi.org/10.1093/bioinformatics/btt593
- Zhang L, Xu P, Cai Y, Ma L, Li S, Li S, Xie W, Song J, Peng L, Yan H, Zou L, Ma Y, Zhang C, Gao Q, Wang J (2017) The draft genome assembly of Rhododendron delavayi Franch. var. delavayi. GigaScience 6 (10): 1-11. https://doi.org/10.1093/gigascience/gix076

Supplementary materials

Suppl. material 1: hDNA laboratory equipment list doi

Authors: Giada Ferrari
Data type: equipment list
Brief description: List of equipment, its purpose and exemplar manufacturer, model and cost (as of 2022) for establishing an hDNA facility.
Download file (13.00 kb)

Suppl. material 2: Case study 2 metadata doi

Authors: Lore Esselens
Data type: Accession information and DNA/sequence quality and quantity statistics
Brief description: Spreadsheet containing sample ID, accession details, collecting data, preservation method, laboratory protocols used and DNA quantity recovery and sequence quality statistics.
Download file (97.06 kb)

Suppl. material 3: Case study 3 accession details doi

Authors: Steven Janssens
Data type: Spreadsheet of accession details
Brief description: Accession details and sample type of the specimens analysed in Case study 3.
Download file (10.44 kb)

Suppl. material 4: Case study 4 accession details doi

Authors: Giada Ferrari
Data type: Spreadsheet of accession details
Brief description: Accession details, links to living and herbarium specimens databases, DNA concentration and library preparation methods.
Download file (7.90 kb)

Suppl. material 5: Case study 4 mapping statistics doi

Authors: Giada Ferrari
Data type: Spreadsheet with mapping statistics for hybridisation capture
Brief description: Read statistics (raw reads, reads mapping to target loci, read clonality, coverage).
Download file (20.72 kb)

Suppl. material 6: Case study 6 sample, DNA and read data description doi

Authors: Gontran Sonet
Data type: Table with descriptive data
Brief description: Sample information, DNA extracts evaluation and DNA reads description.
Download file (17.19 kb)