Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Insect Science

Museums are biobanks: unlocking the genetic potential of the three billion specimens in the world's biological collections

David K Yeates[1], Andreas Zwick[1] and Alexander S Mikheyev[2]

Museums and herbaria represent vast repositories of biological material. Until recently, working with these collections has been difficult, due to the poor condition of historical DNA. However, recent advances in next-generation sequencing technology, and subsequent development of techniques for preparing and sequencing historical DNA, have recently made working with collection specimens an attractive option. Here we describe the unique technical challenges of working with collection specimens, and innovative molecular methods developed to tackle them. We also highlight possible applications of collection specimens, for taxonomy, ecology and evolution. The application of next-generation sequencing methods to museum and herbaria collections is still in its infancy. However, by giving researchers access to billions of specimens across time and space, it holds considerable promise for generating future discoveries across many fields.

Addresses
[1] Australian National Insect Collection, CSIRO National Research Collections Australia, PO Box 1700, Canberra, ACT 2601, Australia
[2] Ecology and Evolution Unit, Okinawa Institute of Science and Technology, 1919-1 Tancha, Onna-son, Kunigami-gun 904-0412, Japan

Corresponding author: Yeates, David K (david.yeates@csiro.au)

Current Opinion in Insect Science 2016, **18**:83–88

This review comes from a themed issue on **Insect phylogenetics**

Edited by **Gregory W Courtney** and **Brian M Wiegmann**

For a complete overview see the Issue and the Editorial

Available online 19th October 2016

http://dx.doi.org/10.1016/j.cois.2016.09.009

Introduction

The molecular revolution has profoundly impacted the biological sciences. In the past decades biobanks and biorepositories have been developed to store tissue or DNA samples appropriate for genetic and genomic research [1], with storage in liquid nitrogen being the 'gold standard'. These biobanks were initiated to store human and model organism accessions, but increasingly their use has been broadened to include a wider range of taxa, and efforts are underway to join them in a virtual network, the Global Genome Biodiversity Network, GGBN [2]. At the time of writing, the GGBN had 49 members around the world storing just over 250,000 tissue samples of only 32,000 species (www.ggbn.org). While this is impressive, these numbers pale in comparison to the estimated 3 billion specimens from 2 million species stored by the world's museums and herbaria [3]. This includes samples of all the nearly 2 million described species and all their synonyms, as well as samples of the about 20,000 species newly described each year [4,5]. Assuming an average genome size of 0.5 Gb, these specimens contain zettabases (10^{21}) of sequence data, on the scale of total available hard drive storage capacity [6]. These museum and herbarium (together termed collections below) samples also span an incredible geographical range across all biomes in all continents and a temporal range extending back prior to the industrial revolution. In addition to their phenotype, what if we could sample the genotype of these specimens?

Increasingly we can do this, with careful extra work. Standard museum and herbarium specimens are often stored at or near room temperature, either air dried or in preservative liquids such as ethanol and formalin. Earlier experimental protocols employing Sanger sequencing technology established that genetic sequences from museum and herbarium specimens were possible to obtain, but success was patchy and limited to genes found in high copy number, such as those from cellular organelles [7]. Similar results were obtained from ancient DNA where genetic sequences were recovered from environmental samples [8]. These early successes highlighted the enormous potential of old DNA samples to answer compelling biological questions, especially about the evolutionary history of extinct species and the nature and trajectory of biological change through time [9–12].

In recent years high throughput sequencing (HTS) technology has greatly expanded and synergised the genetic and genomic potential of biological collections. This is largely because degraded DNA in collection samples is a much more tractable starting point for HTS than previous sequencing technologies, producing greater data yields and the assembly of sequences from a greater variety of genes [13,14••]. The power of new sequencing technology to unlock the genetic and genomic potential of museum and herbarium specimens is so great that it has blurred the distinction between biobank and collection, especially when DNA sequence is the target data. Obviously, there are cases where high-quality biobank tissues are

essential, such as for studies of RNA and complete genome assemblies. Conversely, molecular studies linked to biodiversity benefit more from the taxonomic breadth of biological collections than from a limited number of high quality samples in biobanks. Putting special applications aside, this review focuses on the landscape of opportunities natural history collections offer if looked upon as vast storehouses of genomic DNA, and on important sequencing strategies that open up this rapidly developing field.

Challenges posed by museum specimens

Museum and herbarium specimens pose a number of unique challenges, which require the development of novel molecular and analytic approaches to dealing with them. Below we outline some considerations and proposed solutions.

Damage to the specimen

By definition, museum specimens are irreplaceable, but DNA extraction often results in damage to the specimens. Fortunately, a number of approaches suitable for invertebrates and vertebrates minimize damage to the specimen, while producing adequate DNA yields [15•,16–18].

Fragmentation

The DNA of museum and herbarium specimens will almost certainly be fragmented by a number of processes that begin after death such as DNA hydrolysis through nucleases in the body itself, chemicals used as killing and/or fixing agents, preservatives such as ethanol and formalin [19], and chemicals used to protect against pest attack in the collection such as dichlorvos [20]. Fragmentation is generally not an obstacle for HTS methods because they require short lengths of DNA template. However, extremely short fragments may not carry enough information to be useful, and may need to be filtered out either bioinformatically, or ideally during library construction.

Contamination

Many museum samples contain not just endogenous DNA, but also DNA from bacterial, fungal and other contaminants that have grown in the sample post-mortem. In addition, there is possible contamination from other material that was stored together in the same tray, or vial, or that was brought into contact during specimen handling. Contamination is a major problem, because amplification by polymerase chain reaction (PCR) can bias towards longer, more intact fragments resulting in the overrepresentation of non-endogenous DNA. At best, this consumes sequencing capacity, which can increase the expense of sequencing by over an order of magnitude. At worst, contamination can yield erroneous data. As a result, when working with museum specimens, it is worth following best practices developed for the study of human specimens, such as decontamination with UV-light and physical separation between areas used for DNA extraction and amplification [21].

DNA degradation

DNA breaks down over time, which causes a range of miscoding lesions and can lead to erroneous sequence reads. Depurination, especially in guanosine residues, leads to strand breaks, and deamination of cytosine residues into uracil also occurs [22,23]. Both depurination and deamination can lead to GC→AT sequencing errors. In addition, interstrand cross-linking may occur post-mortem, particularly in formalin-fixed specimens, preventing polymerase bypass, and blocking DNA denaturation [24].

Sequencing strategies

The sequencing strategies are outlined in Figure 1 focus on preparing sequencing libraries from the DNA extract, and, optionally, enriching them for endogenous DNA and evenly targeted sequences.

Direct sequencing of museum samples

While it is certainly possible to sequence museum samples directly, this may not be cost-effective for large numbers of samples, or for organisms with large genomes. In addition, with the exception of the strategy outlined in Figure 1b′, direct sequencing does not eliminate contaminant DNA, which may substantially waste sequencing capacity. However, since unbiased genome representation is often lost during enrichment, direct sequencing may be the best approach for low-input samples. PCR-free libraries are a solution for direct sequencing, which greatly minimize the risk of contamination [25••]. Alternatively, extremely low-coverage whole genome shotgun sequencing ('genome skimming') permits the sequencing of highly abundant DNA (e.g. ribosomal genes, mitochondrial and chloroplast genomes) for a large number of samples, sufficient for many phylogenetic questions [15•,26].

Targeted reduced genome representation

These methods involve the selective capture of genomic regions prior to NGS [27]. Sequence capture methods are technically demanding, require construction of libraries prior to hybridization, and do not scale well [28•]. However, they are cost-effective when dealing with large numbers of samples.

Hybrid enrichment

This method involves hybridizing genomic DNA to DNA probes or 'baits' and then washing away the non-target DNA (e.g. [29–31]). The resulting enriched DNA can be sequenced using various HTS platforms. Bait design requires some knowledge of the target genome, so may require a transcriptome or genome sequence within or adjacent to the target group. The standard probe designs usually work for closely related species with 10–15% divergence, but baits can be designed around targets with