DNA sequencing, meta-barcoding and applications in entomology and taxonomy – a beginner's guide

The basic science underpinning molecular-based approaches is still ultimately reliant on classic taxonomy

Global insect declines are putting ecosystem services at risk [1]. As any conservationist would doubtless concur, without a system in place to accurately measure and contextualise losses, geographic trends and species migrations, a comprehensive understanding of the extent of such apparent declines can be misleading [2,3]. One critical issue is that our ability to monitor and record these declines is highly dependent on taxonomic expertise [4]. Consequently, there is a growing acknowledgement of the breadth and depth of data needed, and the decreasing number of entomological taxonomists [5–7]. Of course, these trends are not occurring in isolation; there has been a growing trend in taxonomic research employing molecular methods – eschewing morphological identification in preference for those that determine species based on DNA.

DNA sequencing in particular has become a central component of the modern diagnostic toolbox. DNA barcoding – a subset of sequencing technologies – involves comparing the genetic code of a part of the animal genome that appears in the same place in every species but, owing to a universally occurring mutation rate, differs very slightly between species [8]. Owing to the ability to compare a single unknown specimen against many potential species in a single assay, and standardised protocols that allow transparent and objective comparison of specimen identifications between laboratories, barcoding is becoming more and more popular [9]. Despite these advantages, the time-consuming processes of extracting DNA and conducting sequencing reactions on individual specimens have mostly limited DNA barcoding for specimen identification to life stages

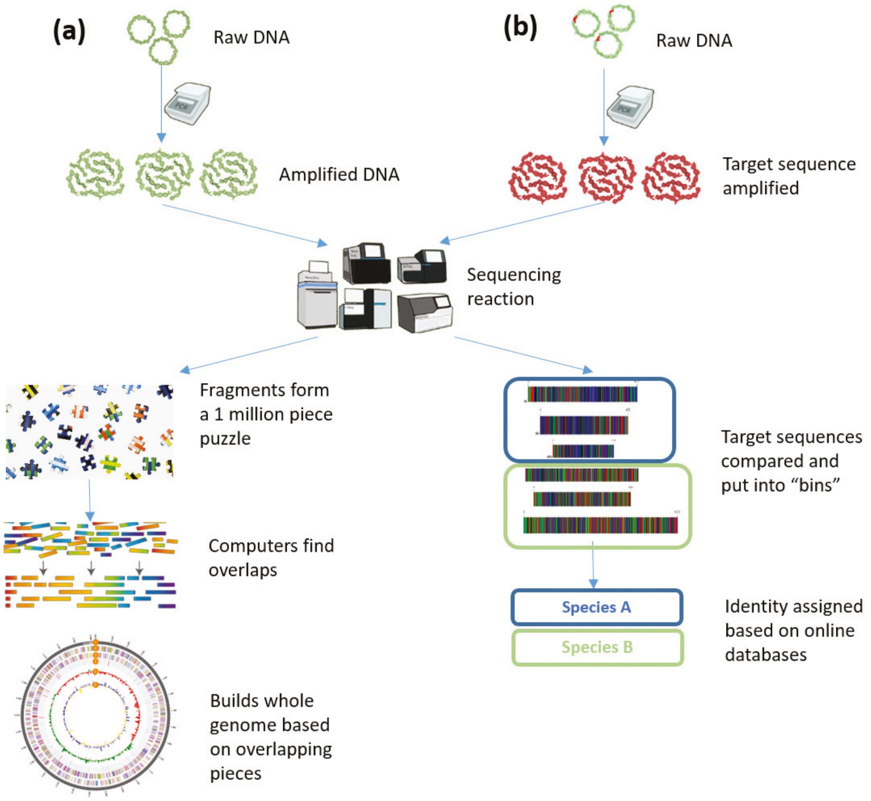
(a)
Raw DNA
Amplified DNA
Fragments form a 1 million piece puzzle
Computers find overlaps
Builds whole genome based on overlapping pieces
(b)
Raw DNA
Target sequence amplified
Sequencing reaction
Target sequences compared and put into "bins"
Species A
Species B
Identity assigned based on online databases

Figure 1. Sequencing approaches take DNA samples extracted from insects, make millions of copies of (a) random fragments, which are then paired together based on overlapping sections for genome sequencing or (b) copies of targeted parts of the genome (a barcode) that are compared based on dissimilarity between fragments, then compared to databases to describe the community composition.

Philip Donkersley
Lancaster Environment Centre,
Lancaster University,
Lancaster, LA1 4YQ, United Kingdom
donkersleyp@gmail.com

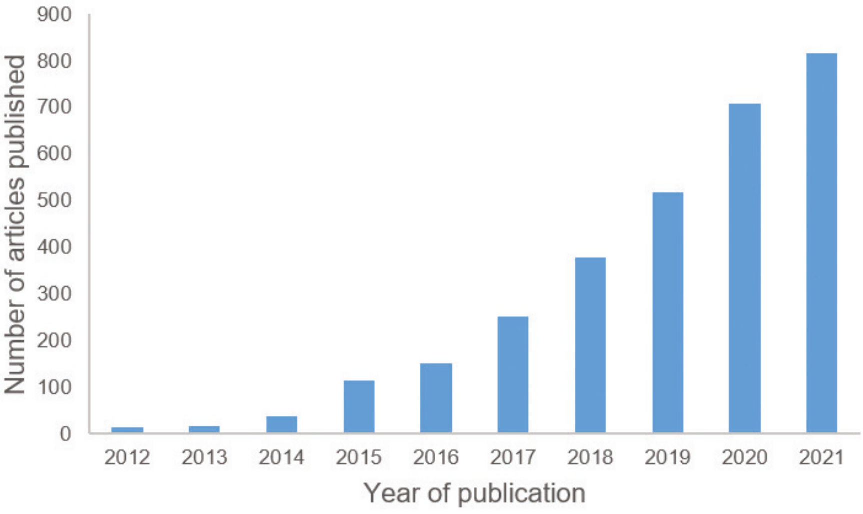

Figure 2. Metabarcoding in the literature. Published articles obtained from Scopus, Crossref, and PubMed databases on 09 Nov. 2021 for all metabarcoding studies.

where a taxonomic key may not be available or important diagnostic structures are degraded or missing [10].

Large numbers of samples are part and parcel of ecological monitoring, the reality of which has so far limited the use of DNA-based approaches; that is, until the recent advent of high-throughput sequencing (HTS) technologies. HTS has allowed DNA barcode-based identification to be conducted relatively cheaply and on a huge number of samples [11]. Thousands of reactions can be done at the same time to produce a huge number of barcode sequences [12], which can in turn be used for genome sequencing (Figure 1a) or "metabarcoding" (simultaneous comparison of different insect species in a mixed sample; Figure 1b).

Due to downstream computer analyses of these sequences, many thousands of individuals can thus be identified rapidly based on comparison to publicly available databases like NCBI Genbank (https://www.ncbi.nlm.nih.gov/genbank/statistics/) and the Barcode of Life Database (https://ibol.org/resources/natural-history-collections/). The speed and breadth of biodiversity surveying possible through metabarcoding means it has increasingly been employed across numerous fields of applied ecology [13,14]. The number of papers published that use a metabarcoding approach has been growing nearly exponentially since 2012 (Figure 2), whilst the cost of sequencing has plummeted in that time [15].

Uptake of molecular tools is about more than just cost; it also encompasses important aspects such as the ease of use, accuracy, reproducibility, up-front investment of training and equipment and compatibility within existing policy frameworks, some or all of which may have hindered widespread collaboration between so-called 'classic' and molecular taxonomists [9]. Fortunately, many excellent resources like the following published papers [12,14–17] and training workshops (*e.g.* Edinburgh Genomics; https://genomics.ed.ac.uk/services/introduction-metagenomic-data-analysis, or The Earlham Institute; https://www.earlham.ac.uk/microbial-analysis) already exist for bridging this gap, but sometimes they miss out on the basics in terms of molecular technological theory and practice.

As for the practical aspects of conducting an ecological survey using metabarcoding approaches, when designing such a survey, the same constraints related to sampling method and collection scheduling that determine the explanatory power of a classic sampling programme apply [18]. Appropriate sampling tool choice is of course crucial, for example: pan trap, blue-vein trap, malaise trap for flying insects; pitfall trap, suction sampler or soil cores for ground or subterranean insects; and various combinations for aquatic invertebrates, *etc*. Despite the high number of sequences that metabarcoding can generate, these sequences do not represent an "individual" animal in the sample pool, and there is some argument over their statistical significance as "pseudoreplicates" [19,20]. Consequently, inadequate sampling and hence insufficient sample replication within a population cannot be "fixed" by intensive sequencing of a small gene pool.

Following sample collection in the field, bulk samples are typically homogenised back in the lab into a "paste" from which DNA can be extracted. Before this stage, various options are open, *e.g.*: preliminary subsampling of trap samples to a taxon of interest (*e.g.* bees), or "bulking up" sample replicates from sites to reduce sequencing costs. After DNA extraction, the next decision is which barcode to use.

Despite current widespread usage for assessing *intra- and inter*specific genetic variation, there is actually a wide choice of different barcodes. Each barcode is limited, with advantages and disadvantages in terms of the level of variation detectable. Barcoding regions represent highly conserved regions, most often found within the mitochondrial genome. The most popularly employed barcode regions are the Cytochrome Oxidase I (COI) and Cytochrome Oxidase II (COII) genes. Similarly, there are structural mitochondrial genes like 12S, 16S, 18S and 28S as well as various inter-gene spacer regions (*e.g.* ITS2), which are commonly employed in metabarcoding studies of insects [9,15,21,22]. In terms of selecting these genes or genetic regions, different studies may involve selection of different sections of the gene concerned (*e.g.* COI has at least three regions commonly used in metabarcoding; Figure 3). Importantly, results from a survey that uses one barcode (*e.g.* COI) will be subtly different from another study in the same region using a different barcode (*e.g.* ITS2), meaning they cannot be accurately compared.

For sure, appropriate selection of a barcode or taxonomic marker is a critical step, since all downstream species detection and identification will necessarily rely on how conserved this marker is across taxa, and hence the discriminatory power of the nucleotide variation contained within it [18].

In terms of a broad survey of the insect genome and its variance, difficulties arise due to a lack of universal cover across insects by any one barcode region, either across different taxa (Figure 3) or different regions (Figure 4). Although COI is arguably fast becoming the barcode of choice for animals, it has been shown to have limited applicability with certain taxa, such as the Tephritidae, due to its lack of interspecific variation [23]. Multi-locus barcoding, where multiple barcoding

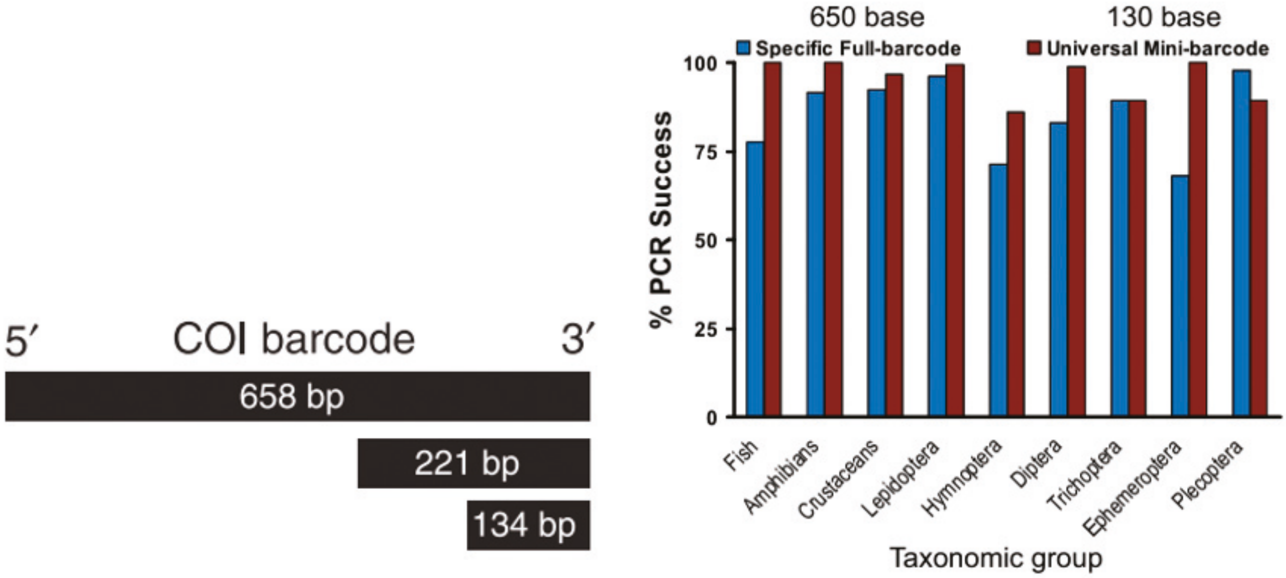

Figure 3. Cytochrome oxidase I gene has three different barcode sections, longer sections offer greater differentiation between species, but produce fewer reads making them less effective on bigger samples. Data from Meusnier *et al.* [30] (Creative Commons CC BY 2.0).

regions are sequenced from the same sample, have the advantage of broader and more accurate species identification, but will also double the financial cost of any sequencing endeavour, and make comparison between studies ultimately even more complicated [24].

Even after considering the cost of sequencing, various manufacturers offer different prices, ways of doing the reactions and the types of experiment they are best adapted to. The diversity of sequencing equipment has grown substantially over the past decade; the market is mostly dominated by two companies: Illumina (https://www.illumina.com/) and PacBio (https://www.pacb.com/). The former employs technology focused on reducing “price per base” sequencing costs and number of reads per sequencing reaction (typically making them the machine of choice for metabarcoding). The latter offers the longest continuous reads of any machine available currently, the company boasting the ability to read an organism’s entire genome in one go, hence eliminating the need for post-sequencing assembly, thereby making these machines more suited to *de novo* genome sequencing. Other small companies produce similar equipment, although most notable of all, Oxford Nanopore has produced a “portable sequencing machine” called the MinION (https://nanoporetech.com/products/minion).

Most sequencing machines use a combination of thermal cycling, DNA polymerase enzymes, spectroscopic lasers and photoreactive DNA markers to replicate the target genome or DNA sequence. From the colour of light given off by the DNA as it replicates the DNA fragment, we can tell what bases are being used, and therefore what the DNA sequence of this new copy of DNA is. The MinION on the other hand uses a passive sequencing approach, rather than a biochemical reaction. This means that it does not replicate the DNA fragments but instead it passes the DNA through a partially-permeable membrane, recording what nucleotides are passing through the membrane in real time.

These machines can work under field conditions, with limited power supply and minimal sample preparation. To date, the MinION has been employed in published studies of microbial ecology using a 16S metabarcoding approach [25], whilst for metabarcoding studies of invertebrates, those involving the COI barcode are just now appearing as preprint manuscripts [26].

In terms of analysing the results obtained from metabarcoding approaches, bioinformatics, the computerised process of converting the sequences into useful information (usually an assembled genome or an ecological community framework) is, despite its complexity, now widely used. These approaches are called pipelines, because rather than being a simple “put in raw data, get out results” system, the pipelines perform a series of different operations using the results from the previous part of the pipeline. They have many additional functions, including eliminating messy or poor-quality reads from the database, producing scores for the “goodness of fit” onto metabarcoding databases, as well as translating sequencing machine data files into formats that can be analysed on statistical software, like R. The diversity of programs is fortunately small, and most of the commonly-used pipelines (*e.g.* QIIME2; https://qiime2.org/) come with extensive instructional aids. The degree of computer literacy required however, often makes hiring a specialist bioinformatics technician a necessity.

The end result of metabarcoding, ideally, is a complete account of all species in the sampled ecosystem, including ones either difficult to identify accurately under a microscope or extremely locally rare species. Yet, despite the advances and hope of greater things to come in terms of better elucidating functional ecology, the methodology and general approach are far from perfect. New species will always need to be identified initially using microscopy, ensuring the need for maintenance of type specimens and the role of natural history museums [27,28]. Despite ever shrinking costs of sequencing, limited science funding in regions critically understudied for biodiversity will limit the impact of the metabarcoding approach in those regions [29]. Furthermore, because of the relative infancy of the technology in the eyes of policy makers, standards and guidelines around its use are still evolving and validated protocols do not yet exist. For example, metabarcoding data cannot yet be used for Water Framework Directive

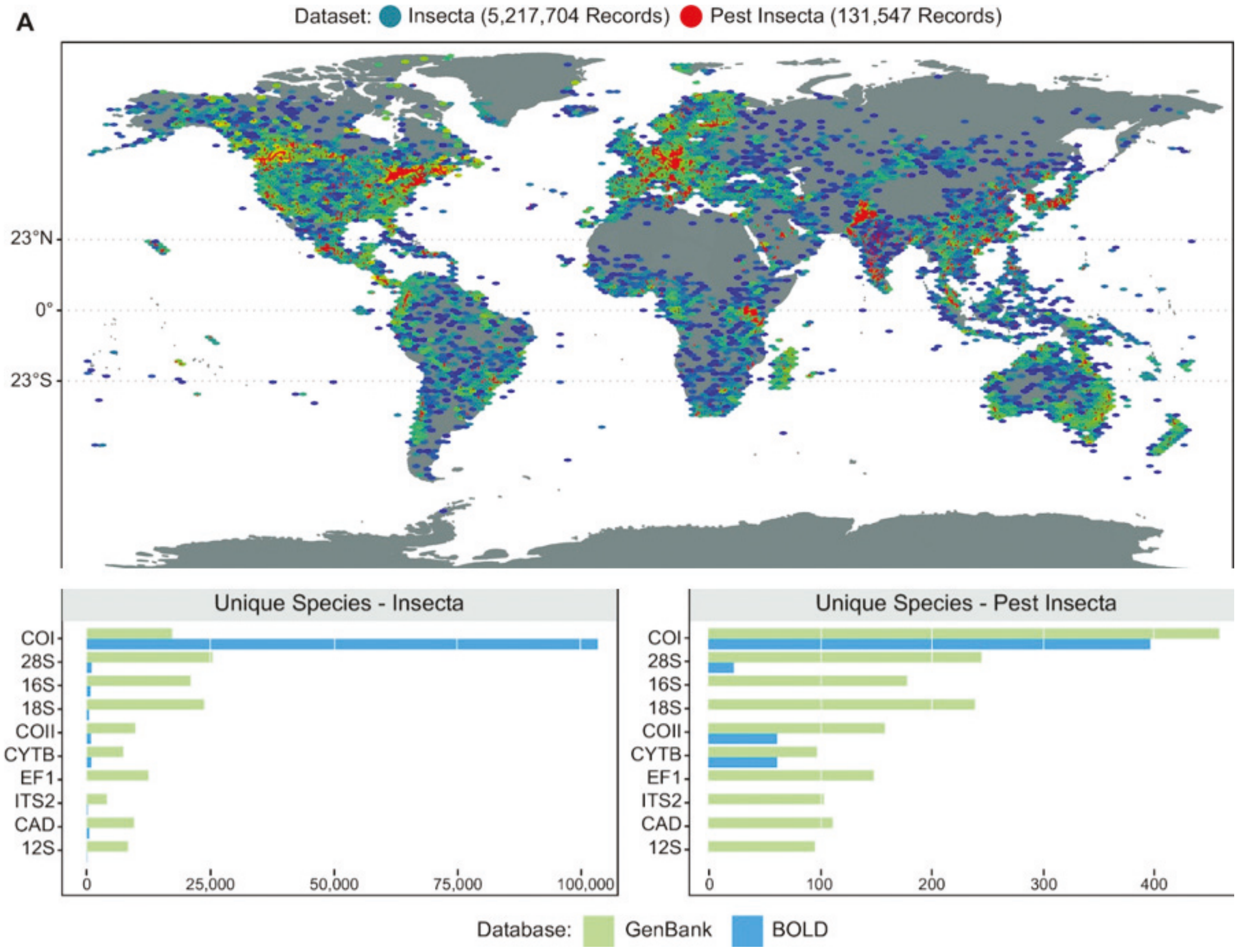

Figure 4. (a) Global distribution of DNA barcodes in lic reference databases. (b) Distribution of records a nique species within major public databases for the 10 barcode markers with the most reference inforn n for Insecta. Data from Piper *et al.* [9] (Creative Con s CC BY).

(WFD) monitoring of benthic macroinvertebrates. This piece of legislation instead only accepts microscope-based community identification as a record of site biodiversity. Finally, as mentioned earlier, comparison between two studies using different barcoding regions is potentially flawed, meaning that although there is a huge number of studies already in existence, we are still far away from an accurate global account of insect biodiversity.

Across Europe, several excellent invertebrate monitoring schemes combine classic microscope-based identification with metabarcoding approaches to supplement the breadth and depth of their data. These include the UK Pollinator Monitoring Scheme (PoMS; https://ukpoms.org.uk/) and Diversity of Insects in Nature protected Areas (DINA; https://www.dina-insektenforschung.de/insekten-monitoring-evk?lang=en) in Germany. The affordability of metabarcoding approaches has substantially improved over the past decade, as has the robustness of reference databases and data-processing pipelines. Consequently, we appear to be on the cusp of a revolution in both the technology and its application, in effect a fortunate paradigm shift in insect monitoring towards metabarcoding approaches. Having said that, the basic science underpinning molecular-based approaches is still ultimately reliant on classic taxonomy – in other words, you can't assign a barcode to a species without the physical type sample!

Acknowledgements

The author receives funding from an ESRC Impact Acceleration Account fund at Lancaster University. I would also like to thank Dr Richard Harrington for inviting me to write this article, and for the RES Genomics SIG meeting that originally inspired preparing this material.

References

1. Sánchez-Bayo, F. *et al.* (2019) *Biological Conservation* **232**, 8–27.
2. Daskalova, G.N. *et al.* (2021) *Insect Conservation and Diversity* **14**, 149–154.
3. Harvey, J.A. *et al.* (2020) *Nature Ecology & Evolution* **4**, 174–176.
4. Deng, J. *et al.* (2019) *Insect Conservation and Diversity* **12**, 18–28.
5. Hopkins, G.W. *et al.* (2002) *Animal Conservation* **5**, 245–249.
6. Cheesman, O.D. *et al.* (2007) *Proceedings of the Royal Entomological Society's 23rd Symposium*; ISBN 9781845932541.
7. Orr, M.C.C. *et al.* (2020) *Megataxa* **1,** 19–27.
8. Ratnasingham, S. *et al.* (2013) *PLoS One* **8**, e66213.
9. Piper, A.M. *et al.* (2019) *Gigascience* **8**, giz092.
10. Krehenwinkel, H. *et al.* (2019) *Genes (Basel)* **10**, 858.
11. Alberdi, A. *et al.* (2018) *Methods in Ecology and Evolution* **9**, 134–147.
12. Taberlet, P. *et al.* (2012) *Molecular Ecology* **21**, 2045–2050.
13. Tedersoo, L. *et al.* (2019) *Molecular Ecology Resources* **19**, 47–76.
14. Deiner, K. *et al.* (2017) *Molecular Ecology* **26**, 5872–5895.
15. Liu, M. *et al.* (2020) *Ecological Entomology* **45,** 373–385.
16. Staats, M. *et al.* (2016) *Analytical and Bioanalytical Chemistry* **408**, 4615–4630
17. Elbrecht, V. *et al.* (2021) *PeerJ* **9**, e12177.
18. Zinger, L. *et al.* (2019) *Molecular Ecology* **28**, 1857–1862.
19. Bush, A. *et al.* (2019) *Frontiers in Ecology and Evolution* **7**, 434.
20. Paulson, J.N. *et al.* (2013) *R Packages*.
21. Baird, D.J. *et al.* (2012) *Molecular Ecology* **21**, 2039–2044.
22. Morinière, J. *et al.* (2019) *Molecular Ecology Resources* **19**, 900–928.
23. Jiang, F. *et al.* (2014) *Molecular Ecology Resources* **14**, 1114–1128.
24. Zhang, G.K. *et al.* (2018) *Evolutionary Applications* **11**, 1901–1914.
25. Kai, S. *et al.* (2019) *FEBS Open Biology* **9**, 548–557.
26. Abeynayake, S.W. *et al.* (2021) *Genes (Basel)* **12**, 1138.
27. Call, E. *et al.* (2021) *Insect Systematics and Diversity* **5**, 6.
28. Salvador, R.B. *et al.* (2020) *Oecologia* **192**, 641–646.
29. Simmons, B.I. *et al.* (2019) *Ecology and Evolution* **9**, 3678–3680.
30. Meusnier, I. *et al.* (2008) *BMC Genomics* **9**, 214.