The LINNEAN SOCIETY of London BIOLOGICAL Journal of the Linnean Society

Biological Journal of the Linnean Society, 2016, **117**, 5–10.

COMMENT

Collections-based research in the genomic era

SVEN BUERKI[1]* and WILLIAM J. BAKER[2] FLS

[1]*Department of Life Sciences, Natural History Museum, Cromwell Road, London, SW7 5BD, UK*
[2]*Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AB, UK*

Received 22 October 2015; accepted for publication 22 October 2015

Biological collections are at the front line of biodiversity research, informing taxonomy, evolution, conservation and sustainable livelihoods. In April 2014, we organised a meeting at the Linnean Society (UK) discussing the impact of next-generation sequencing (NGS) methods on collections-based research. Here, we explore the main themes of this meeting and outline the incredible potential of NGS to reinvent collections-based research. Among the many opportunities at the interface of genomics and collections, we focus specifically on (1) the genomic characterisation of biological collections, (2) the enhancement and development of DNA-based identification, (3) the tree of life and (4) interdisciplinary research addressing the most pressing environmental challenges of our times. Across the world, biological collections are at risk, primarily due to declining funding and shifts in scientific fashions. We encourage all users of collections to embrace the genomic era, not only because of the unparalleled scientific potential that it presents, but also because new cross-disciplinary synergies will reinvigorate and secure the collections for future generations. © 2015 The Linnean Society of London, *Biological Journal of the Linnean Society*, 2016, **117**, 5–10.

ADDITIONAL KEYWORDS: museomics – next-generation sequencing – phylogenomics – phylogeny – taxonomy – tree of life.

INTRODUCTION

Collections-based research, traditionally conducted in Natural History Museums, Botanical Gardens and Zoos, has constantly re-invented itself in response to changing methods and technologies. Biological collections, such as museum or herbarium specimens, living organisms and DNA samples, are at the front line of biodiversity research, informing taxonomy, evolution, conservation and sustainable livelihoods. The emergence of next-generation sequencing (NGS) methods has dramatically changed the face of biology, and is leading to the rapid generation of massive amounts of DNA sequence data at reduced costs (e.g. Doyle, 2013). This leap forward in sequencing technology has been made possible by a combination of nanotechnology, advanced imaging methods and high-powered computing. NGS is having a similar impact on collections-based research as the molecular systematic revolution of the 1990s. It is therefore vital that researchers working with collections develop a full understanding of the benefits and pitfalls of these technologies, and learn to exploit them to the full. Consequently, as the genomic era dawns collections-based researchers and curators must adapt and respond to the new opportunities that these technologies present. This challenge was the focus of a meeting at the Linnean Society of London (UK) on 2–3 April 2014. The eleven contributions published in this special issue encapsulate the outcome of this fruitful event, and are wide-ranging in their use of collections in ancient DNA research (Linderholm, 2016), evolution (Kidner *et al.*, 2016), bioinformatics (Vieira *et al.*, 2016), phylogenomics (Dodsworth *et al.*, 2016; Heyduk *et al.*, 2016), systematics (Dentinger *et al.*, 2016; Zedane *et al.*, 2016) and museomics (Bailey *et al.*, 2016; Bakker *et al.*, 2016; Besnard *et al.*, 2016; Timmermans *et al.*, 2016). In this short opinion, we touch on the primary themes explored in this meeting while sharing some of our own views on the subject, especially in relation to the incredible potential of NGS to reinvent collections-based research.

*Corresponding author. E-mail: s.buerki@nhm.ac.uk

THE GENOMIC ERA – A REVOLUTION FOR BIOLOGICAL COLLECTIONS

The application of traditional DNA sequencing methods to museum and herbarium collections has been hindered by their state of preservation, especially in the case of older material, which typically yields degraded DNA of unusable quality (e.g. Devey *et al.*, 2013). However, NGS can overcome this obstacle because the technology actually requires fragmented DNA (e.g. Bakker *et al.*, 2016). In the case of plants, NGS techniques have even succeeded with herbarium specimens preserved in alcohol in the field prior to drying (the so-called Schweinfurth technique, widely used in the tropics where field drying is difficult; Bridson & Forman, 2010), which until now were very difficult to sequence due to their low DNA yields (Särkinen *et al.*, 2012). In this special issue, seven milestone publications provide detailed laboratory and bioinformatic protocols to reconstruct organellar genomes from historical collections across the animal (horseshoe bats: Bailey *et al.*, 2016; insects: Timmermans *et al.*, 2016; pigeons: Besnard *et al.*, 2016), fungal (Dentinger *et al.*, 2016) and plant (angiosperms: Bakker *et al.*, 2016; palms: Heyduk *et al.*, 2016; olive family: Zedane *et al.*, 2016) kingdoms. These pioneering studies provide us with a toolbox for exploiting an almost limitless genetic resource that has been out of reach until now. Our own institutions illustrate the breathtaking potential of these approaches. Amassed over the past 300 years, London's Natural History Museum and the Royal Botanic Gardens, Kew house around 90 million specimens, including >34 million insects and arachnids, 29 million other animals, 12 million plants and 1.25 million fungi. Most of the known biodiversity of Earth is represented within these two institutions that are separated by just a short 20 min ride on the London Underground. The scale of the genomic opportunities that they now present are hard to comprehend, but now, more than ever before, they are poised to be deployed in genomic research addressing some the biggest science challenges facing humankind today. We expand on some of these challenges here.

GENOMIC CHARACTERISATION OF BIOLOGICAL COLLECTIONS

The genomic toolbox presented in this special issue provides a unique opportunity to add another dimension to biological collections by characterising them from a molecular standpoint. In the not too distant future, we anticipate that Natural History Museums and Botanic Gardens will automate high-throughput sequencing of their collections, delivering the resulting data via portals centralising all available collections information (e.g. Wen *et al.*, 2015 for more details on databases related to biodiversity). Many major institutions are actively developing large-scale specimen digitisation programmes (see Smith & Blagoderov, 2012 and references therein). To future-proof these efforts, the progressive step of sampling the collections for genomic analyses could be included in the workflow. Though an upfront additional cost, this extra stage would create future efficiencies, removing the need for another collections upheaval for genomic sampling, and might create opportunities for pilot studies aimed at developing high-throughput workflows for the sequencing at a massive taxonomic scale. The potential applications of large-scale genomic characterisation of major biological collections are limited only by the imagination of the users. We elaborate on a few of the more obvious applications below, but many others exist, for example in microevolution (studies below the species level), comparative molecular evolution, and the genomics of specimen traits, which could potentially be mined from digitised collections databases (cf. Pyron, 2015).

IDENTIFICATION BARCODES

The genomic characterisation of collections will also greatly benefit the rapid identification of species using the DNA barcoding approach (see Barcode of Life Data System; Ratnasingham & Hebert, 2007). In addition, the vast amount of DNA data obtained from large-scale sequencing of collections would also enable development of novel DNA barcodes specific to groups which cannot be discriminated by the widely used barcode regions (e.g. several lineages of monocots such as Pandanaceae; Buerki *et al.*, 2012). The release of large amounts of molecular data to identify species would boost taxonomy, especially in the case of organisms exhibiting cryptic morphological features (e.g. fungi, Dentinger *et al.*, 2016; beetles, Triponez *et al.*, 2011; parasitoid-wasps: Kenyon *et al.*, 2015) or those difficult-to-preserve groups for which identification is very difficult. As an example of the latter case, the economically and ecologically important plant genus *Piper* (*c.* 2000 species, belonging to the peppercorn family) has been proposed as a model to study evolution, chemical ecology, and trophic interactions by Dyer & Palmer (2004). However, this genus lacks a global taxonomic treatment, partly because dried material is very difficult to use, which hinders species delimitations and the construction of keys. As a result, it is impossible to accurately associate the occurrence of any chemical

compound with species and therefore infer their position within trophic network, a problem exacerbated by the size of the genus. The sequencing of nuclear and chloroplast DNA barcodes would considerably boost research on this economically and ecologically important genus by providing a framework to delimit and identify species.

INFERRING THE TREE OF LIFE

Type collections as a means to infer the tree of life

Despite ongoing global efforts, our understanding of the tree of life remains incomplete due to inadequacy of data sampling from the majority of its branches (e.g. Hinchliff & Smith, 2014; Hinchliff *et al.*, 2015). The challenge presented by this sampling gap has until recently seemed insurmountable due to the time and funding required to obtain new samples from those missing branches. However, NGS methods now allow us to exploit the sampling efforts made by the generations of field collectors who have deposited their specimens in museums and herbaria over many centuries. If researchers seize this opportunity, NGS methods will bring about a golden age of tree of life research. The scale and potential of the data opportunity may also persuade reluctant curators that sampling from those most precious specimens, the types, can be justified. In taxonomy, types are gold standard specimens that determine the correct application of nomenclature (e.g. Knapp *et al.*, 2004 and references therein). The use of types in collections-based research is therefore central to ensuring accurate species identification by default. As taxonomists, we deplore the inadequate documentation of evidence supporting species identification in many phylogenetic studies, which not only devalues such research, but also undermines the reliability of public nucleotide databases (e.g. Vilgalys, 2003; Valkiunas *et al.*, 2008; Marucci, La Rosa & Pozio, 2010). In the current biodiversity crisis, accurate species identification is a fundamental research requirement, or can even be seen as a kind of critical infrastructure. In this context, the tree of life will be instrumental in establishing a genomic and phylogenetic framework that can be used by the biodiversity community to rapidly identify species and if necessary describe them. Another advantage of using type specimens is the increasing availability of images and data from types through public digitisation initiatives (e.g. Global Plants Initiative; Ryan, 2013), paving the way for seamless links between databases of nucleotide records and of the specimens from which those nucleotides were derived.

Towards a total-evidence tree of life ... including extinct species

Before the molecular systematics revolution, phylogenetic inferences where based on morphological characters and therefore tended to include both extant and extinct species (i.e. total-evidence phylogenetics; Donoghue *et al.*, 1989). However, such an approach rapidly declined due to the difficulties of (1) obtaining DNA sequences for extinct taxa and (2) combining morphological and DNA data (Scotland, Olmstead & Bennett, 2003). However, NGS coupled with ancient DNA will rejuvenate the inclusion of extinct taxa into phylogenetic frameworks as shown by the study in this issue on the extinct plant genus *Hesperelaea* (Oleaceae; Zedane *et al.*, 2016). This approach still has some limitations with most of the successfully analyzed species being extinct during the Holocene (e.g. Brace *et al.*, 2015; but see Miller *et al.*, 2008 for the sequencing of the nuclear genome of woolly mammoth from the Pleistocene). However, in the case where DNA sequencing cannot be performed from fossils, methods have been recently developed that allow the integration of morphological and molecular data for both fossils and extant species (see Barreda *et al.*, 2015 and reference therein). Overall, the inclusion of extinct lineages into the tree of life should significantly improve topological and branch-length estimation (e.g. Wiens *et al.*, 2010), and will also allow more accurate biogeographical reconstructions (e.g. following an approach similar to Meseguer *et al.*, 2015). In this context, we would encourage further collaborations at the frontier between paleontology and neontology to infer the tree of life.

COLLECTIONS-BASED RESEARCH AND ENVIRONMENTAL CHALLENGES

There is a growing body of evidence showing that we have entered a new era of mass extinction (the Anthropocene), driven by human activities similar in both rate and magnitude to the last big five extinctions (Dirzo *et al.*, 2014). For instance, a recent study estimated that the extinction rate of vertebrates is currently 114 times higher than the normal rate (Ceballos *et al.*, 2015). Collections-based research could play a major role in minimising the effect of climate change and habitat destruction on biodiversity. For instance, an accurately inferred tree of life together with DNA barcodes based on collections could be used as references to support rapid biodiversity surveys based on environmental DNA (e.g. from soil, water or faeces; Taberlet *et al.*, 2012) or the inference of the network of trophic interactions to assess ecosystem functioning and sustainability (see

Hrček & Godfray, 2015 and references therein). One of the major obstacles in these fields of research is the lack of a reference library for species identifications. The gold mine of genetic resources available in herbaria and natural history museums can alleviate this issue, but ecologists, collections-based researchers and conservation managers will have to work in synergy to reach this goal.

The addition of knowledge on species spatial distributions – obtained through programs of collections digitisation (e.g. the Global Biodiversity Information Facility platform: http://www.gbif.org) – to the tree of life can shed light on the means by which biota were shaped by using a phylogenetic community approach coupled with species modeling (e.g. Buerki *et al.*, 2015 for Madagascar). This latter information can then provide the basis to further predict how biota will respond to ongoing anthropogenic activities. However, the full potential of the use of collections for conservation and climate change research cannot be fully realised until high-quality data sets are conveniently accessible to researchers. This requires that higher priority be placed on digitising the holdings most useful to support this type of research (e.g. whole-biota studies, time series, records of intensively sampled common taxa).

CONCLUDING REMARKS

The papers in this special issue and others cited here shine a spotlight on the genomic opportunities within collections that are now coming within reach. Across the world, biological collections face many threats, due to failing financial support from governments and universities, the decline of traditional expertise, and unfavorable shifts in academic expectations. This short-sightedness is paradoxical given that the global genomic repository that is the world's biological collections is now on the brink of giving up its genetic secrets. All custodians and users of biological collections must now embrace genomics, not only because of the unparalleled scientific potential it presents, but also because cross-disciplinary synergies will reinvigorate and secure the collections for future generations. The research illustrated here is just the beginning of a new genomic collections era.

ACKNOWLEDGEMENTS

We are very grateful to the Linnean Society of London, especially Tom Simpson, for providing the support and ideal conditions that made the 'Collections-based research in the genomic era' conference so successful. We would like to thank all the conference speakers and contributors to the special issue. We acknowledge the financial support of the Linnean Society, the Centre for Ecology and Evolution and BMC Evolutionary Biology, without which the conference and this volume would not have been possible. Finally, we are very grateful to the editor in-chief, Prof. John Allen, for his support in putting together the special issue.

REFERENCES

Bailey SE, Mao X, Struebig M, Tsagkogeorga G, Csorba G, Heaney LR, Sedlock J, Stanley W, Rouillard J-M, Rossiter SJ. 2016. The use of museum samples for large-scale sequence capture: a study of congeneric horseshoe bats (family Rhinolophidae). *Biological Journal of Linnean Society* **117:** 58–70.

Bakker FT, Lei D, Yu J, Mohammadin S, Wei Z, van de Kerke S, Gravendeel B, Nieuwenhuis M, Staats M, Alquezar-Planas DE, Holmer R. 2016. Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biological Journal of Linnean Society* **117:** 33–43.

Barreda VD, Palazzesi L, Tellería MC, Olivero EB, Raine JI, Forest F. 2015. Early evolution of the angiosperm clade Asteraceae in the Cretaceous of Antarctica. *Proceedings of the National Academy of Sciences of the United States of America* **112:** 10989–10994.

Besnard G, Bertrand JAM, Delahaie B, Bourgeois YXC, Lhuillier E, Thébaud C. 2016. Valuing museum specimens: high-throughput DNA sequencing on historical collections of New Guinea crowned pigeons (*Goura*). *Biological Journal of Linnean Society* **117:** 71–82.

Brace S, Turvey ST, Weksler M, Hoogland MLP, Barnes I. 2015. Unexpected evolutionary diversity in a recently extinct Caribbean mammal radiation. *Proceedings of the Royal Society B* **282:** 20142371.

Bridson D, Forman L. 2010. *The herbarium handbook*. Kew: Kew Publishing.

Buerki S, Callmander MW, Devey DS, Chappell L, Gallaher T, Munzinger J, Haevermans T, Forest F. 2012. Straightening out the screw-pines: a first step in understanding phylogenetic relationships within Pandanaceae. *Taxon* **61:** 1010–1020.

Buerki S, Callmander MW, Bachman S, Moat J, Labat J-N, Forest F. 2015. Incorporating evolutionary history into conservation planning in biodiversity hotspots. *Philosophical Transactions of the Royal Society, Series B* **370:** 20140014.

Ceballos G, Ehrlich PR, Barnosky AD, García A, Pringle RM, Palmer TM. 2015. Accelerated modern human–induced species losses: entering the sixth mass extinction. *Science Advances* **1:** e1400253.

Dentinger BTM, Gaya E, O'Brien H, Suz LM, Lachlan R, Díaz-Valderrama JR, Koch RA, Aime MC. 2016.

Tales from the crypt: genome mining from fungarium specimens improves resolution of the mushroom tree of life. *Biological Journal of Linnean Society* **117:** 11–32.

Devey DS, Forest F, Rakotonasolo F, Ma P, Dentinger BTM, Buerki S. 2013. A snapshot of extinction in action: the decline and imminent demise of the endemic *Eligmocarpus* Capuron (Caesalpinioideae, Leguminosae) serves as an example of the fragility of Madagascan ecosystems. *South African Journal of Botany* **89:** 273–280.

Dirzo R, Young H, Galetti M, Ceballos G, Isaac N, Collen B. 2014. Defaunation in the Anthropocene. *Science* **345:** 401–406.

Dodsworth S, Chase MW, Särkinen T, Knapp S, Leitch AR. 2016. Using genomic repeats for phylogenomics: a case study in wild tomatoes (*Solanum* section *Lycopersicon*: Solanaceae). *Biological Journal of Linnean Society* **117:** 96–105.

Donoghue MJ, Doyle JA, Gauthier J, Kluge AG, Rowe T. 1989. The importance of fossils in phylogeny reconstruction. *Annual Review of Ecology and Systematics* **20:** 431–460.

Doyle JJ. 2013. The promise of genomics for a 'next generation' of advances in higher-level legume molecular systematics. *South African Journal of Botany* **89:** 10–18.

Dyer LE, Palmer ADN. 2004. *Piper: a model genus for studies of phytochemistry, ecology, and evolution*. New York, NY: Kluwer Academics.

Heyduk K, Trapnell DW, Barrett CF, Leebens-Mack J. 2016. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biological Journal of Linnean Society* **117:** 106–120.

Hinchliff CE, Smith SA. 2014. Some limitations of public sequence data for phylogenetic inference (in plants). *PLoS ONE* **9:** e98986.

Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazis R, Gude K, Hibbett DS, Katz LA, Laughinghouse HD IV, McTavish EJ, Midford PE, Owen CL, Ree RH, Rees JA, Soltis DE, Williams T, Cranston KA. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences of the United States of America* **112:** 12764–12769.

Hrček J, Godfray HC. 2015. What do molecular methods bring to host–parasitoid food webs? *Trends in Parasitology* **31:** 30–35.

Kenyon SG, Buerki S, Hansson C, Alvarez N, Benrey B. 2015. Uncovering cryptic parasitoid diversity in *Horismenus* (Chalcidoidea, Eulophidae). *PLoS ONE* **10:** e0136063.

Kidner C, Groover A, Thomas DC, Emelianova K, Soliz-Gamboa C, Lens F. 2016. First steps in studying the origins of secondary woodiness in *Begonia* (Begoniaceae): combining anatomy, phylogenetics, and stem transcriptomics. *Biological Journal of Linnean Society* **117:** 121–138.

Knapp S, Lamas G, Lughadha EN, Novarino G. 2004. Stability or stasis in the names of organisms: the evolving codes of nomenclature. *Philosophical Transactions of the Royal Society, serie B* **359:** 611–622.

Linderholm A. 2016. Ancient DNA: the next generation – chapter and verse. *Biological Journal of Linnean Society* **117:** 150–160.

Marucci G, La Rosa G, Pozio E. 2010. Incorrect sequencing and taxon misidentification: an example in the Trichinella genus. *Journal of Helminthology* **84:** 336–339.

Meseguer AS, Lobo JM, Ree R, Beerling DJ, Sanmartin I. 2015. Integrating fossils, phylogenies, and niche models into biogeography to reveal ancient evolutionary history: the case of *Hypericum* (Hypericaceae). *Systematic Biology* **64:** 215–232.

Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, Tomsho LP, Packard MD, Zhao F, Sher A, Tikhonov A, Raney B, Patterson N, Lindblad-Toh K, Lander ES, Knight JR, Irzyk GP, Fredrikson KM, Harkins TT, Sheridan S, Pringle T, Schuster SC. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456:** 387–390.

Pyron RA. 2015. Post-molecular systematics and the future of phylogenetics. *Trends in Ecology & Evolution* **30:** 384–389.

Ratnasingham S, Hebert PDN. 2007. BOLD: the barcode of life data system (www.barcodinglife.org). *Molecular Ecology Notes* **7:** 355–364.

Ryan D. 2013. The global plants onitiative celebrates its achievements and plans for the future. *Taxon* **62:** 417–418.

Särkinen T, Staats M, Richardson JE, Cowan RS, Bakker FT. 2012. How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLoS ONE* **7:** e43808.

Scotland RW, Olmstead RG, Bennett JR. 2003. Phylogeny reconstruction: the role of morphology. *Systematic Biology* **52:** 539–548.

Smith VS, Blagoderov V. 2012. Bringing collections out of the dark. *ZooKeys* **209:** 1–6.

Taberlet P, Coissac E, Hajibabaei M, Riesberg LH. 2012. Environmental DNA. *Molecular Ecology* **21:** 1789–1793.

Timmermans MJTN, Viberg C, Martin G, Hopkins K, Vogler AP. 2016. Rapid assembly of taxonomically validated mitochondrial genomes from historical insect collections. *Biological Journal of Linnean Society* **117:** 83–95.

Triponez Y, Buerki S, Borer M, Naisbit R, Rahier M, Alvarez N. 2011. Discordances between phylogenetic and morphological patterns of alpine leaf beetles attest to intricate history of lineages in postglacial Europe. *Molecular Ecology* **20:** 2442–2463.

Valkiunas G, Atkinson CT, Bensch S, Sehga RNM, Ricklefs RE. 2008. Parasite misidentifications in GenBank: how to minimize their number? *Trends in Parasitology* **24:** 247–248.

Vieira FG, Lassalle F, Korneliussen TS, Fumagalli M. 2016. Improving the estimation of genetic distances from Next-Generation Sequencing data. *Biological Journal of Linnean Society* **117:** 139–149.

Vilgalys R. 2003. Taxonomic misidentification in public DNA databases. *New Phytologist* **160:** 4–5.

Wen J, Ickert-Bond SM, Appelhans MS, Dorr LJ, Funk VA. 2015. Collections-based systematics: opportunities and

outlook for 2050. *Journal of Systematics and Evolution* **53:** 477–488.

Wiens JJ, Kuczynski CA, Townsend T, Reeder TW, Mulcahy DG, Sites JW Jr. 2010. Combining phylogenomics and fossils in higher-level squamate reptile phylogeny: molecular data change the placement of fossil taxa. *Systematic Biology* **59:** 674–688.

Zedane L, Hong-Wa C, Murienne J, Jeziorski C, Baldwin BG, Besnard G. 2016. Museomics illuminate the history of an extinct, paleoendemic plant lineage (*Hesperelaea*, Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico. *Biological Journal of Linnean Society* **117:** 44–57.