



**Cite this article:** Hollingsworth PM, Li D-Z, van der Bank M, Twyford AD. 2016 Telling plant species apart with DNA: from barcodes to genomes. *Phil. Trans. R. Soc. B* **371**: 20150338.  
<http://dx.doi.org/10.1098/rstb.2015.0338>

Accepted: 1 June 2016

One contribution of 16 to a theme issue 'From DNA barcodes to biomes'.

**Subject Areas:**

ecology, evolution, taxonomy and systematics, plant science

**Keywords:**

plant DNA barcoding, next-generation sequencing, genome skimming, species discrimination, hybrid baits

**Author for correspondence:**

Peter M. Hollingsworth  
e-mail: [p.hollingsworth@rbge.org.uk](mailto:p.hollingsworth@rbge.org.uk)

# Telling plant species apart with DNA: from barcodes to genomes

Peter M. Hollingsworth<sup>1</sup>, De-Zhu Li<sup>2</sup>, Michelle van der Bank<sup>3</sup>  
and Alex D. Twyford<sup>4</sup>

<sup>1</sup>Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, UK

<sup>2</sup>Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, 132 Lanhei Road, Heilongtan, Kunming, Yunnan 650201, People's Republic of China

<sup>3</sup>Department of Botany and Plant Biotechnology, University of Johannesburg, Auckland park, Johannesburg PO Box 524, South Africa

<sup>4</sup>Ashworth Laboratories, Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK

 PMH, 0000-0003-0602-0654; ADT, 0000-0002-8746-6617

Land plants underpin a multitude of ecosystem functions, support human livelihoods and represent a critically important component of terrestrial biodiversity—yet many tens of thousands of species await discovery, and plant identification remains a substantial challenge, especially where material is juvenile, fragmented or processed. In this opinion article, we tackle two main topics. Firstly, we provide a short summary of the strengths and limitations of plant DNA barcoding for addressing these issues. Secondly, we discuss options for enhancing current plant barcodes, focusing on increasing discriminatory power via either gene capture of nuclear markers or genome skimming. The former has the advantage of establishing a defined set of target loci maximizing efficiency of sequencing effort, data storage and analysis. The challenge is developing a probe set for large numbers of nuclear markers that works over sufficient phylogenetic breadth. Genome skimming has the advantage of using existing protocols and being backward compatible with existing barcodes; and the depth of sequence coverage can be increased as sequencing costs fall. Its non-targeted nature does, however, present a major informatics challenge for upscaling to large sample sets.

This article is part of the themed issue 'From DNA barcodes to biomes'.

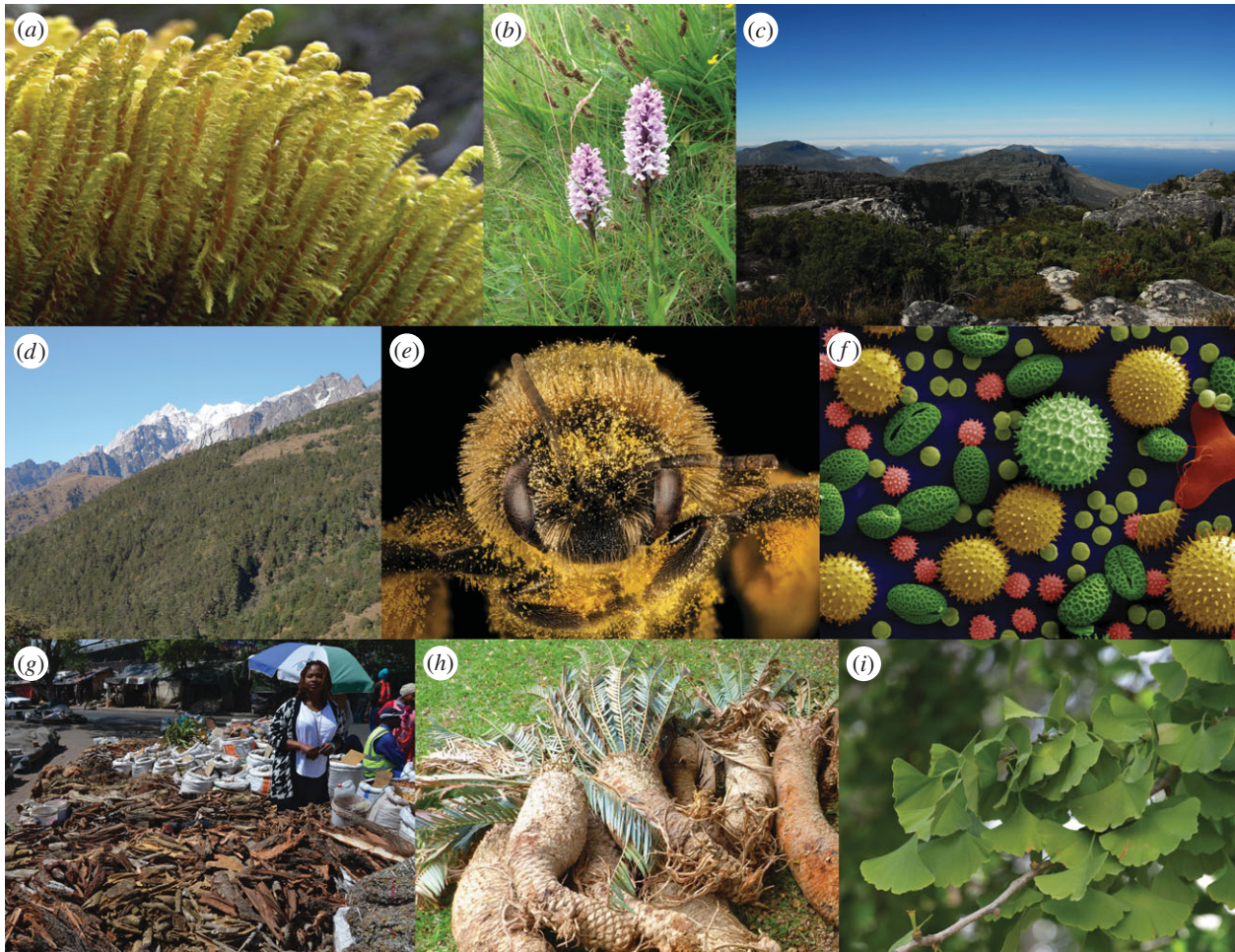
## 1. Introduction

Despite centuries of taxonomic effort, the characterization of plant species diversity remains a substantial and important challenge. Although plants are undoubtedly well understood compared to mega-diverse groups like insects, recent estimates suggest that around 70 000 flowering-plant species await discovery [1]. Beyond finding new species, existing taxonomic accounts need reconciling and updating, and there is also the wider practical challenge of assigning unidentified specimens to known species. This latter point is particularly pertinent where the available material is sub-optimal (e.g. juvenile, fragmented, processed) or where available levels of taxonomic expertise are low.

DNA barcoding involves the standardized use of one or a few DNA regions to tell species apart [2]. In this paper, we summarize the extent to which DNA barcoding of plants [3] is providing practical progress to address these challenges and also explore the opportunities presented by the ongoing development of new sequencing technologies.

## 2. Standard plant barcodes

There is no single plant barcode that matches the universality and resolving power of Cytochrome Oxidase (C01) in animals [2]. Most specimen-based



**Figure 1.** Example uses of DNA barcoding. (a) Species discovery in the bryophyte *Herbertus* (Herbertaceae). Image: David Genney, (b) first complete national DNA barcode database, for the flora of Wales. Image: Alex Twyford, (c) floristic barcoding of the Cape Flora. Image: Olivier Maurin, (d) DNA barcoding the flora of China. Image: De-Zhu Li, (e) pollen identification and the study of pollen movement. Image: USGS Bee Inventory and Monitoring Lab, (f) species identification of historical mixed pollen samples. Image: Dartmouth Electron Microscope Facility, (g) a stand selling plant products in Johannesburg. Image: Zandisile Shongwe, (h) confiscated illegal *Encephalartos* (Zamiaceae), Image: Eastern Cape Department of Economic Development, Environmental Affairs and Tourism, (i) identification of plant compounds (here extract from *Ginkgo biloba*) in herbal supplements. Image: Juan Carlos Lopez Almansa.

plant-barcoding studies use one or a few plastid regions (e.g. the protein coding ‘core barcodes’ *rbcL* and *matK*, and the non-coding spacer *trnH-psbA*) and the internal transcribed spacer (ITS) regions of nuclear ribosomal DNA (ITS—either its entirety or just the ITS2 region) [4–7]. Plant studies focusing on mixed templates and/or degraded DNAs (e.g. environmental samples) typically use the P6 loop of the plastid *trnL* intron, whose short length and conserved primer sequences make it particularly amenable to amplification and short-read sequencing via next-generation sequencing (NGS) technologies [8,9].

In many animal groups, the close concordance of species with barcode sequence clusters enables the semi-automated quantification of species diversity [10,11]. However, plant-plastid and ribosomal-DNA barcodes typically have lower discriminatory power [12] and do not lead to tight clustering of conspecifics separated by clear discontinuities from other species in sequence space. Instead, there is typically a graded continuum of intra- and interspecific distances, with barcodes commonly shared among related species [12]. There are two main implications of this. Firstly, standard plant barcodes are best suited to being molecular augmentations to existing classifications, rather than having the resolving power to act in a stand-alone fashion to define a species-level framework. Secondly, in using plant barcodes, attention should be given at the outset to ensuring a match between the resolving power of the technique, and the

information that is required from the study. Examples of the range of studies plant barcodes are being used for are given below.

### (a) Species discovery

Plant barcodes are typically used in an integrative fashion with other information for detecting new taxa. In some studies, unexpected sequence divergence has led to re-examination of morphological/ecological variation, which has then resulted in formal recognition of new species [13]. In other cases, morphological or ecological variants have been the trigger for generating sequence data to establish whether there is supporting genetic evidence for recognizing different taxa [14]. Species discovery has involved the full spectrum of species from relatively small and/or character-poor groups like bryophytes (figure 1a) through to conspicuous ecologically/culturally important trees, and in a small number of cases, the nucleotide variants themselves have been formalized into the species descriptions (e.g. [15,16]).

### (b) Vegetation and floristic surveys

Geographically restricted floristic assemblages represent an inherently lower discrimination challenge for plant barcodes, as the closest relatives of many taxa may be absent from the area. Floristic barcoding projects have been completed at a



**Table 1.** Levels of species discrimination from floristic barcoding studies at different scales and levels of floristic complexity.

study type	study location	no. species	markers	species discrimination (%)	references
tropical trees, forest plot	16-ha plot, northeast Puerto Rico	143	<i>rbcl</i> , <i>matK</i> , <i>trnH-psbA</i>	93	[17]
tropical trees, forest plot	50-ha plot, Cameroon	272	<i>rbcl</i> , <i>matK</i> , <i>trnH-psbA</i>	71–88	[18]
nature reserve	348-ha, Ontario, Canada	436	<i>rbcl</i> , <i>matK</i> , <i>trnH-psbA</i>	95	[19]
nature reserve	1133-ha, Guangdong, China	417	<i>rbcl</i> , ITS2	65	[20]
local flora	20 000-ha Churchill, Manitoba, Canada	312	<i>rbcl</i> , <i>matK</i> , ITS2	69	[21]
national flora	2 m-ha, Wales, UK	1041	<i>rbcl</i> , <i>matK</i>	69–75	[22]
(large) regional flora	Canadian arctic	490	<i>rbcl</i> , <i>matK</i>	56	[23]

range of scales from plot-level studies of tropical trees [17,18], nature reserves [19,20], local flora [21] and small countries (figure 1*b*, [22]). Ongoing large-scale barcoding projects include steps to complete the barcoding of the flora of South Africa and an ambitious multi-institute project to barcode the flora of China (figure 1*c,d*). Not surprisingly, there is large variation in the percentage of species discriminated (table 1), and this is strongly affected by the geographical scale of study and the complexity of the flora [20,22,24]. Although many other factors are at play, the larger the scale of the study, and the greater the number of species-rich genera, the lower the discrimination success.

Moving back in time from contemporary floristic barcoding, several studies have used the *trnL* intron P6 loop to reconstruct historical vegetation types based on environmental sequencing from frozen sediments dating back thousands of years (e.g. [25–27]). Although the small size of the P6 loop (which makes it so well-suited for recovery from ancient samples) inevitably constrains its resolving power at the species level, the approach does provide a standard scalable approach for broad-brush identification, which can increase resolution beyond that of morphological palynology in some plant groups [26].

### (c) Ecological forensics

Floristic barcoding datasets provide a foundation for studies of ecological processes. Conventional identification of plants from individual tissue types/juvenile life stages is usually difficult as the seedlings, roots, seeds and pollen and other gametophytes of many species can appear similar. If the material has been processed in one way or another (e.g. been digested), the difficulties of identification are exacerbated. Thus, as with paleobarcoding, even barcode datasets with imperfect species resolution can still provide knowledge gains. For instance, Kartzinel *et al.* [28] barcoded faecal samples from African herbivores and showed clear dietary niche partitioning even among similar coexisting species. Likewise, Kesanakurti *et al.* [29] used barcode data to show strong spatial structuring of plant roots in the absence of corresponding above-ground structuring. The field of pollen barcoding is growing rapidly, and even modest increases in discriminatory power beyond morphological identification (figure 1*e,f*) hold great promise to enhance understanding of the dynamics and consequences of pollination and pollen movement [30–32].

### (d) Identification to support regulatory enforcement

Reliable identification of plant material by regulatory/enforcement authorities is a widespread need, including

identification of pests, pathogens and invasive species to inform control [33,34], detecting protected species being illegally traded (figure 1*g*, [35,36]), through to identifying food or herbal medicine labelling errors/fraud (figure 1*h*, [37]).

Although some applications require species-level resolution, many do not. For instance, useful insights into the composition of food and drink can be obtained at the level of them containing ‘something other than what is on the label’, and Stoeckle *et al.* [38] showed that about one-third of herbal teas contained plant species beyond those listed. Likewise, many studies have deployed DNA barcoding approaches to assess the plant components of herbal medicines and dietary supplements, and evidence of market substitution/adulteration is not uncommon [39–41]. For instance, Little [42] found evidence that 3/37 Ginkgo herbal supplements contained fillers with no detectable Ginkgo DNA (figure 1*i*), and Kumar *et al.* [43] showed evidence for widespread mislabelling of Bala herbal products in market samples.

### (e) DNA barcoding and community phylogenetics

The differing levels of variability among standard plant barcode regions means that commonly deployed markers (e.g. *rbcl*, *matK*, *trnH-psbA* and ITS) can provide resolution at different phylogenetic levels, which has facilitated studies of community phylogenies [17,44], comparative biology and phylogenetic diversity. Shapcott *et al.* [45] used plastid barcodes to identify priority areas for conservation in Australian rainforests based on both species richness and phylogenetic diversity (involving the identification of areas containing more phylogenetic diversity than would be expected based on species richness alone). Using floristic phylogenies in a rather different manner, Saslis-Lagoudakis *et al.* [46] capitalized on barcode datasets for the floras of the Cape of South Africa, Nepal and New Zealand to study the phylogenetic distribution of plants used in traditional medicine. They showed significant phylogenetic clustering of traditionally used medicinal species and highlighted cases where different cultures have exploited the same lineages for bioactive compounds, and noted the predictive capacity of the phylogenies for further screening for bioactives.

## 3. Limitations of standard plant barcodes

Pilot studies, careful project design and an appropriate match of inference to the level of signal in the data are critical to the effective use of standard plant barcodes, and these principles underpin many of the studies described above. This is necessary as the literature is replete with examples of plants sharing

barcodes among related species and numerous cases where uniquely distinguishable species in a genus are the exception not the rule [12]. Thus, uncritical use of plant barcoding may lead to disappointing and/or uninformative results. Beyond this fundamental challenge of restricted/variable discriminatory power, there are additional practical issues such as primer mismatches impacting on the recovery of *matK* barcodes, as well as ongoing different preferences for different barcode regions for different applications that make it difficult to combine reference datasets generated for different purposes or studies [12].

#### 4. Factors influencing the discriminatory power of standard plant barcodes

Various studies have been undertaken to unpick the reasons why plant barcodes are often shared between related species. Obvious drivers include hybridization, groups with slow mutation rates relative to speciation rate, and general challenges in groups showing recent and rapid divergence [12,24,47,48]. Somewhat less obvious is the notion that the limited seed dispersal compared with pollen dispersal in many plant species may act as an intrinsic limitation on the degree to which maternally inherited plastid barcodes are likely to track species boundaries [12,49]. This is attributable to the low intra-specific gene flow of seed-dispersed plastid markers essentially retarding the ability of new variants to spread throughout a species range, and a related increase in the likelihood of successful local interspecific introgression [12,50]. Likewise, selective sweeps acting on the plastid genome combined with hybridization have also been invoked in limiting the resolving power of plastid barcodes—best exemplified by the remarkable case of *Salix*, where 337 individuals from 53 species from 3/5 subgenera across Europe and North America share a barcode haplotype [51,52]. The use of nuclear ITS often increases levels of resolution beyond those of plastid markers but within limits [5,6]. In some groups, multiple copies occur, creating challenges of sequencing and/or interpretation of paralogues, and interspecific barcode sharing either through lack of divergence or hybridization is also not uncommon in ITS datasets [5].

Plant barcoding is at something of a crossroads. On the one hand, there are a multitude of applications that are well suited for the existing resolving power of plant barcodes and continuing these studies, and establishing sample sets to support the reference databases remains a high priority and focus for the plant-barcoding community. On the other hand, given the limitations of discriminatory power of standard barcodes in many plant groups, there is a clear and unambiguous need for improved barcoding protocols.

#### 5. Extending and improving the plant barcode

##### (a) Additional amplicon sequencing

One option to improve species discrimination in plants is supplementing standard plant DNA barcodes with additional loci generated with Sanger Sequencing or via NGS of tagged amplicons [53]. The benefits here are that Sanger Sequencing is inexpensive on a per-individual basis, and that there have been methodological improvements in generating these data (e.g. improved DNA polymerases). While there are some

candidates and improvements in discrimination for individual groups [53,54], most evidence suggests that the gain in species discrimination will be incremental [55,56]. This is particularly the case as barcoding with Sanger Sequencing is limited to organellar and ribosomal loci, as cloning heterozygous nuclear loci is not feasible. Even where Sanger Sequencing gives way to NGS of barcoded amplicons, these approaches are typically constrained to a small set of nuclear loci, and evidence to date suggests sometimes very modest discrimination gains from sequencing  $\sim 10$  nuclear regions as barcodes due to lack of intraspecific coalescence [57,58].

##### (b) Plastid genome sequences

Several authors have argued for having complete plastid gene sets, or indeed complete plastid genomes, as the plant barcode [59,60]. The highly conserved gene order, the absence of recombination and low levels of nucleotide substitution make the plastid the ideal target for comparative analysis across the land plants. In addition, the high-copy number means genomic DNA extracts are enriched for plastids, and thus an easier target than low-copy nuclear genes for sequencing, particularly from degraded samples. Complete (or near-complete) plastid genomes can be obtained by short- or long-ranged PCR enrichment with conserved primers [61,62], direct isolation protocols [63], capture via oligonucleotide probes [64] or genome skimming of genomic DNA [65].

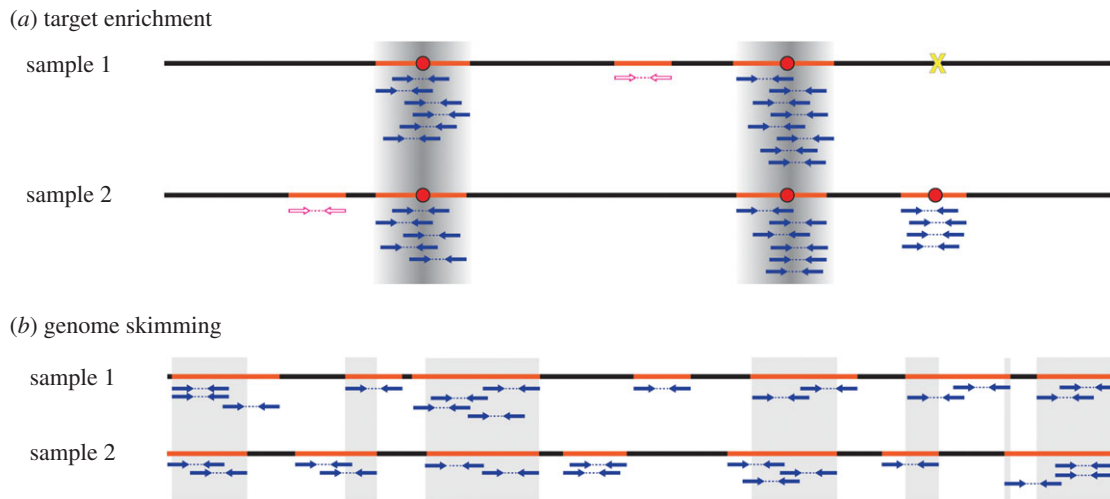
Sequencing the complete plastid genome provides more characters and increases the amount of sequence data by two orders of magnitude (e.g. from approx. 1400 bp for *rbcl* and *matK* to approx. 150 000 bp), and this can provide some increase in species discrimination (e.g. [58]). Use of complete plastid genomes also gets around the problem of different research groups favouring different plastid regions, as the reference database essentially covers all plastid barcodes [66]. Complete plastid genome sequencing fits the requirement of being highly scalable, with reliable automated assemblers (e.g. ORG.asm assembler [67]), annotation [68] and broad-scale alignment [69] possible for all but the most structurally divergent land-plant plastid genomes such as those found in parasitic, mycoheterotrophic or carnivorous taxa [70].

However, complete plastid sequencing does not address the basic challenge that plastid genomes do not necessarily track species boundaries [12,71,72]. Thus, although we envisage the coming few years will see a steep increase in the number of publications using complete plastid genomes as barcodes, the ultimate big gains in resolving power will only come with access to substantial numbers of unlinked nuclear markers. There are two obvious primary routes to do this: target enrichment or genome skimming, with additional possibilities including transcriptome sequencing and RAD-seq.

##### (c) Targeted enrichment

Targeted enrichment includes approaches that use short oligonucleotide probes (baits) to pull down homologous sequences in a genomic DNA extract, with the enriched DNA then subject to NGS (figure 2a, [73]). The approach is highly scalable and well suited for recovery from degraded DNAs [73,74]. The key question here is whether a universal probe set can be developed to capture a large set of homologous loci across all land plants [66].

While it is clear that there are nuclear loci conserved across large groups of plants, such as the 1025 conserved orthologue



**Figure 2.** Comparison between promising genomic barcodes. (a) Target enrichment focuses sequencing reads (blue arrows) on homologous regions of the genome surrounding bait sites (red dots), with many regions with high coverage (dark-grey shading). Samples missing a suitable bait site (yellow cross) are not represented in the data. Off-bait reads (pink open arrows) may be informative, particularly if they map to high-copy ribosomal DNA or organelles. (b) Genome skimming can be used to generate a fragmented nuclear assembly with low sequence coverage. Homologous sequences are a random collection of regions where assemblies overlap (grey boxes).

set loci between tomato and *Arabidopsis* [75], or the 1083 putative orthologues in the genomes of seven angiosperms and a moss species [76], there is no single set of well-curated nuclear genes. However, there are a wealth of resources that could be used to find conserved loci, particularly transcriptomes from the oneKP project ([77] [onekp.com](http://onekp.com)), or the 58 complete plant genomes (<https://phytozome.jgi.doe.gov>). A conserved set of baits designed from these resources could be supplemented with available baits for large and important clades, such as the Compositae [78]. As an important issue is the balance between low-variation universal loci, or more variable loci that can only be retrieved from a subsample of species.

Going from a candidate gene set to an effective hybridization assay is non-trivial. Early studies showed that most conserved *Arabidopsis* loci do not hybridize to tomato baits under stringent conditions [75], probably due to sequence divergence. On the other hand, using lower stringency conditions will capture off-target sequences and paralogues, which will affect downstream applications. Here, the challenge is designing short probes that can effectively bait a single specific target locus. A landmark study in the application of baits to phylogenetically disparate taxa [64] used a suite of 55 000 RNA baits, each 120 bp in length, to capture entire plastids. The short probe length used here would be particularly useful for capturing loci from degraded samples. However, it is unclear whether such a phylogenetically diverse range of land plants could be assayed with a single nuclear gene set. This is a high-priority area for assay development.

#### (d) Genome skimming

A genomic DNA extract typically contains a mix of nuclear and organellar DNA (plastid and mitochondria), and NGS will generate data across the three genomes. At low sequence coverage (e.g. 0.1–10 $\times$ , approx. 1 GB of data), the genome can be ‘skimmed’ [65], allowing the near-complete assembly of the high-copy plastid, mitochondria and ribosomal RNA (figure 2b). There is also the potential to make a highly fragmented nuclear genome assembly.

Genome skimming has great promise for extending the plant barcode, reviewed by Coissac *et al.* [66]. Importantly, genome skimming is scalable and (relatively) cost-effective, and can be used effectively with degraded DNAs from herbarium specimens [79]. At the lower end, benchtop protocols for single insert-size library preparation, such as the Illumina Nextera and TruSeq, can be performed on a small number of samples. For larger applications, library preparation can be automated on robotic liquid handlers such as the Illumina NeoPrep. These libraries can then be multiplexed on a range of sequencing platforms, with the cheapest per-sample-costs with high-output platforms (box 1). Downstream, parts of the data assembly are suited to automation, particularly organelle assembly (e.g. plastids [67,79], mitochondria [85]). In terms of costs, library preparation and low-coverage sequencing can be \$200 per sample when highly multiplexed [66].

A second benefit of genome skimming is that it is both backwards-compatible with the standard plant barcodes, and forwards-compatible with genome sequencing (discussed below) [66]. Genome skims routinely recover plastid barcode loci and ITS, and thus continue to add to the growing reference database of the standard barcoding loci. In terms of compatibility with future genome sequencing approaches, the archived sequence reads that can be reassembled as improved assembly algorithms become available, while archived DNA samples or NGS libraries could be resequenced to provide better coverages as costs decrease [66].

A significant challenge for using genome skimming for DNA barcoding is how to effectively use the nuclear data. Many genome-skimming studies discount the nuclear reads and only assemble the organellar and ribosomal DNA [65,86,87]. While nuclear assemblies are possible using assemblers intended for large diploid genomes (reviewed in [79]), the combined factors of low sequence coverage, short-read lengths and single small DNA insert size means the nuclear assembly will be a near-random collection of fragmented DNA sequences. An assembly from a single-insert library will often have a median contiguous DNA size (N50) of around 5–10 kb, with the largest fragments in the range of 30–120 kb in length (AD Twyford, 2016 unpublished data). To

**Box 1.** Recent developments in NGS platforms.

*Increased output.* The Illumina HiSeq X and HiSeq 4000 sequencers use patterned flow cell technology to generate extremely high output. The HiSeq 4000 generates 750 GB data per run, enough to sequence 90 *Arabidopsis* genomes at 60× coverage. These platforms will greatly reduce the cost of projects that use a large number of short reads (up to 150 bp paired-end), such as genome skims.

*Longer read lengths.* Current long-read sequencers include the Pacific BioSciences real-time sequencer [80] and Oxford NanoPore's MinION [81]. These PCR-free single molecular sequencing platforms generate reads many kilobases in length (PacBio > 10 kb, MinION > 5 kb), with these data widely used to scaffold genomes assembled from inexpensive Illumina data [82]. Their immediate use for barcoding is unclear due to their high error rates and sequencing costs, though proof-of-concept studies suggest that these platforms are promising [83].

*Portable sequencers.* Oxford NanoPore's MinION is the first portable NGS platform. This pocket-sized device allows sequencing to be done anywhere, only requiring a connection to a laptop. Other benefits include the low lease cost and the production of data in real time. Portable genomics has great potential and may enable barcoding in the field. While field-based sequencing has become reality for studying the spread of viruses [84], for field barcoding of plants there will need to be new sample assays that focus the modest sequencing output onto homologous regions.

*In-house genomics.* The high purchase costs and the requirement for specialized laboratory skills have limited NGS platforms to large centralized sequencing hubs. This is likely to change with the release of low-output sequencing platforms intended for small research groups. The most prominent is the Illumina MiniSeq, which costs \$50 000, has a small footprint, and produces 7.5 GB of data overnight. This platform could be extremely useful for barcoding work with amplicons or enriched samples, such as those from hybrid baits. It could also be used for preliminary genomics of challenging samples such as those from degraded tissues.

use this for species discrimination will rely on algorithms that can cope with comparisons among sample sets with highly variable and patchy overlap in the data [66].

**(e) Other technologies**

One of the most popular approaches to access large numbers of nuclear markers is through the sequencing of regions adjacent to restriction-enzyme cut sites, including genotyping by sequencing [88] and restriction site-associated DNA sequencing (RAD [89,90]). These approaches allow thousands of homologous regions to be sequenced across hundreds of individuals, without prior knowledge of the genome sequence. While there are cases where these methods have been informative across species clades (e.g. [91–93]), the lack of conserved cut sites across a very broad taxonomic scope makes them better suited to closely related taxa. While RAD has its benefits, and deserves more thorough testing for species discrimination in individual clades, we do not see this as a primary route for universal barcoding.

Transcriptome sequencing is a widely used tool for the analysis of gene expression, marker discovery and comparative evolution [94]. The main benefit of transcriptomics is that it focuses NGS onto a homologous proportion of the genome, which in this case is also the most highly conserved. However, the requirement for high-quality fresh material, and the tissue-specific nature of the sequences, rules it out for universal barcoding.

**(f) Entire genomes**

The gold standard in genome sequencing are model organisms such as *Drosophila*, *Arabidopsis* and humans, where sequence reads mapped to high-quality reference genomes allow chromosome-level assemblies encompassing most of the genome [95]. There are also many cases where high-quality reference genomes have been assembled *de novo*

from diverse wild organisms [96,97]. While many plant genomes are now publically available, there are major technical and biological hurdles to making whole-genome sequencing scalable and cost-effective. The biggest limitation to *de novo* plant genome assembly are repetitive sequences and the associated large variation and size of plant genomes (plants vary 2000-fold in their genome sizes [98], with a number of groups containing species with giant genomes, e.g. more than 40 GB in *Fritillaria*, [99]). And although there are many other reasons why huge datasets of complete genome sequences are highly desirable, for the particular challenge of species discrimination there would be substantial redundancy in the data.

**6. Concluding remarks**

In this paper, we have outlined the strengths and diverse applications of standard plant barcodes but also noted their limitations. We have summarized some of the exciting future directions made possible by developments in sequencing technologies. However, it is important to qualify this future enthusiasm with a healthy dose of pragmatism. DNA barcoding involves huge sample sets [10]. Part of its success has been based around industrial-scale thinking of laboratory practices and informatics pipelines. The challenges of data editing, quality checking, analysis and storage for standard barcodes are far from trivial, and massively upscaling the depth of data per individual is a huge undertaking. Likewise, although NGS costs continue to fall, the per-sample library preparation costs are still prohibitive in many cases. Large-scale projects involving thousands of samples are underway using genome skimming [66], and the informatics pipelines are progressing rapidly. There are, however, considerable developments and cost reductions required before 'Plant Barcoding 2.0' can be considered truly scaleable and widely adoptable, especially to less well-resourced laboratories. With this in mind, we advocate a



twin-track approach of (i) continued construction of the reference library via large-scale sample sets and careful deployment of standard plant barcodes, while (ii) maintaining and enhancing international collaborative efforts to further develop plant barcode protocols to support the ultimate objective of establishing a workflow with the resolving power to uniquely discriminate the vast majority of the world's land plant species.

**Authors' contributions.** All authors contributed to the development of ideas and information in this paper.

**Competing interests.** We have no competing interests.

## References

- Bebber DP *et al.* 2010 Herbaria are a major frontier for species discovery. *Proc. Natl Acad. Sci. USA* **107**, 22 169–22 171. (doi:10.1073/pnas.1011841108)
- Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003 Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B* **270**, 313–321. (doi:10.1098/rspb.2002.2218)
- Kress JW, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. 2005 Use of DNA barcodes to identify flowering plants. *Proc. Natl Acad. Sci. USA* **102**, 8369–8374. (doi:10.1073/pnas.0503123102)
- Hollingsworth PM *et al.* 2009 A DNA barcode for land plants. *Proc. Natl Acad. Sci. USA* **106**, 12 794–12 797. (doi:10.1073/pnas.0905845106)
- China-Plant-BOL-Group. 2011 Comparative analysis of a large dataset indicates that ITS should be incorporated into the core barcode for seed plants. *Proc. Natl Acad. Sci. USA* **108**, 19 641–19 646. (doi:10.1073/pnas.1104551108)
- Hollingsworth PM. 2011 Refining the DNA barcode for land plants. *Proc. Natl Acad. Sci. USA* **108**, 19 451–19 452. (doi:10.1073/pnas.1116812108)
- Kress WJ, Erickson DL. 2007 A two-locus global DNA barcode for land plants: The coding *rbcl* gene complements the non-coding *trnH-psbA* spacer region. *PLoS ONE* **2**, e508. (doi:10.1371/journal.pone.0000508)
- Taberlet P *et al.* 2007 Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.* **35**, e14. (doi:10.1093/nar/gkl938)
- Valentini A *et al.* 2009 New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trnL* approach. *Mol. Ecol. Resour.* **9**, 51–60. (doi:10.1111/j.1755-0998.2008.02352.x)
- Hebert PDN, Ratnasingham S, Zakharov EV, Telfer AC, Levesque-Beaudin V, Milton MA, Pedersen S, Jannetta P, deWaard JR. 2016 Counting animal species with DNA barcodes: Canadian insects. *Phil. Trans. R. Soc. B* **371**, 20150333. (doi:10.1098/rstb.2015.0333)
- Ratnasingham S, Hebert PDN. 2013 A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS ONE* **8**, e66213. (doi:10.1371/journal.pone.0066213)
- Hollingsworth PM, Graham SW, Little DP. 2011 Choosing and using a plant DNA barcode. *PLoS ONE* **6**, e19254. (doi:10.1371/journal.pone.0019254)
- Liu J, Moeller M, Gao L-M, Zhang D-Q, Li D-Z. 2011 DNA barcoding for the discrimination of Eurasian yews (*Taxus L.*, Taxaceae) and the discovery of cryptic species. *Mol. Ecol. Resour.* **11**, 89–100. (doi:10.1111/j.1755-0998.2010.02907.x)
- Liu Y-J, Newmaster SG, Wu X-J, Liu Y, Ragupathy S, Motley T, Long C-L. 2013 *Pinellia hunanensis* (Araceae), a new species supported by morphometric analysis and DNA barcoding. *Phytotaxa* **130**, 1–13. (doi:10.11646/phytotaxa.130.1.1)
- Bell D, Long DG, Forrest AD, Hollingsworth ML, Blom HH, Hollingsworth PM. 2012 DNA barcoding of European *Herbertus* (Marchantiopsida, Herbertaceae) and the discovery and description of a new species. *Mol. Ecol. Resour.* **12**, 36–47. (doi:10.1111/j.1755-0998.2011.03053.x)
- Filipowicz N, Nee M, Renner S. 2012 Description and molecular diagnosis of a new species of *Brunfelsia* (Solanaceae) from the Bolivian and Argentinean Andes. *Phytokeys* **10**, 83–94. (doi:10.3897/phytokeys.10.2558)
- Kress WJ, Erickson DL, Swenson NG, Thompson J, Uriarte M, Zimmerman JK. 2010 Advances in the use of DNA barcodes to build a community phylogeny for tropical trees in a Puerto Rican forest dynamics plot. *PLoS ONE* **5**, e15409. (doi:10.1371/journal.pone.0015409)
- Parmentier I, Duminiel J, Kuzmina M, Philippe M, Thomas DW, Kenfack D, Chuyong GB, Cruaud C, Hardy OJ. 2013 How effective are DNA barcodes in the identification of African rainforest trees? *PLoS ONE* **8**, e54921. (doi:10.1371/journal.pone.0054921)
- Burgess KS, Fazekas AJ, Kesanakurti PR, Graham SW, Husband BC, Newmaster SG, Percy DM, Hajibabaei M, Barrett SCH. 2011 Discriminating plant species in a local temperate flora using the *rbcl*+*matK* DNA barcode. *Methods Ecol. Evol.* **2**, 333–340. (doi:10.1111/j.2041-210X.2011.00092.x)
- Liu J, Yan H-F, Newmaster SG, Pei N, Ragupathy S, Ge X-J. 2015 The use of DNA barcoding as a tool for the conservation biogeography of subtropical forests in China. *Divers. Distrib.* **21**, 188–199. (doi:10.1111/ddi.12276)
- Kuzmina ML, Johnson KL, Barron HR, Hebert PDN. 2012 Identification of the vascular plants of Churchill, Manitoba, using a DNA barcode library. *BMC Ecol.* **12**, 25. (doi:10.1186/1472-6785-12-25)
- de Vere N *et al.* 2012 DNA barcoding the native flowering plants and conifers of Wales. *PLoS ONE* **7**, e37945. (doi:10.1371/journal.pone.0037945)
- Saarela J, Sokoloff P, Gillespie L, Consaul L, Bull R. 2013 DNA barcoding the Canadian Arctic flora: core plastid barcodes (*rbcl*+*matK*) for 490 vascular plant species. *PLoS ONE* **8**, e77982. (doi:10.1371/journal.pone.0077982)
- Clement WL, Donoghue MJ. 2012 Barcoding success as a function of phylogenetic relatedness in *Viburnum*, a clade of woody angiosperms. *BMC Evol. Biol.* **12**, 73. (doi:10.1186/1471-2148-12-73)
- Jorgensen T *et al.* 2012 Islands in the ice: detecting past vegetation on Greenlandic nunataks using historical records and sedimentary ancient DNA Meta-barcoding. *Mol. Ecol.* **21**, 1980–1988. (doi:10.1111/j.1365-294X.2011.05278.x)
- Sønsteby JH *et al.* 2010 Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Mol. Ecol. Resour.* **10**, 1009–1018. (doi:10.1111/j.1755-0998.2010.02855.x)
- Willerslev E *et al.* 2014 Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature* **506**, 47–51. (doi:10.1038/nature12921)
- Kartzinel TR, Chen PA, Coverdale TC, Erickson DL, Kress WJ, Kuzmina ML, Rubenstein DI, Wang W, Pringle RM. 2015 DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proc. Natl Acad. Sci. USA* **112**, 8019–8024. (doi:10.1073/pnas.1503283112)
- Kesanakurti PR, Fazekas AJ, Burgess KS, Percy DM, Newmaster SG, Graham SW, Barrett SCH, Hajibabaei M, Husband BC. 2011 Spatial patterns of plant diversity below-ground as revealed by DNA barcoding. *Mol. Ecol.* **20**, 1289–1302. (doi:10.1111/j.1365-294X.2010.04989.x)
- Kraaijeveld K, Weger LA, Ventayol García M, Buermans H, Frank J, Hiemstra PS. 2015 Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA

- sequencing. *Mol. Ecol. Resour.* **15**, 8–16. (doi:10.1111/1755-0998.12288)
31. Richardson RT, Lin C-H, Sponsler DB, Quijia JO, Goodell K, Johnson RM. 2015 Application of ITS2 metabarcoding to determine the provenance of pollen collected by honey bees in an agroecosystem. *Appl. Plant Sci.* **3**, 1400066. (doi:10.3732/apps.1400066)
  32. Sickel W, Ankenbrand MJ, Grimmer G, Holzschuh A, Härtel S, Lanzen J, Steffan-Dewenter I, Keller A. 2015 Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC Ecol.* **15**, 1–9. (doi:10.1186/s12898-015-0051-y)
  33. Cross HB, Lowe AJ, Gurgel CFD. 2011 DNA barcoding of invasive species. In *Fifty years of invasion ecology: the legacy of Charles Elton* (ed. DM Richardson), pp. 289–299. Oxford, UK: Wiley-Blackwell.
  34. Hoveka LN, van der Bank M, Boatwright JS, Bezeng BS, Yessoufou K. 2016 The noncoding trnH-psbA spacer, as an effective DNA barcode for aquatic freshwater plants, reveals prohibited invasive species in aquarium trade in South Africa. *South Afr. J. Bot.* **102**, 208–216. (doi:10.1016/j.sajb.2015.06.014)
  35. Aubriot X, Lowry PP, Cruaud C, Couloux A, Haevermans T. 2013 DNA barcoding in a biodiversity hot spot: potential value for the identification of Malagasy *Euphorbia* L. listed in CITES Appendices I and II. *Mol. Ecol. Resour.* **13**, 57–65. (doi:10.1111/1755-0998.12028)
  36. Hartvig I, Czako M, Kjaer ED, Nielsen LR, Theilade I. 2015 The use of DNA barcoding in identification and conservation of Rosewood (*Dalbergia* spp.). *PLoS ONE* **10**, e138231. (doi:10.1371/journal.pone.0138231)
  37. Wallace LJ, Boilard SMAL, Eagle SHC, Spall JL, Shokralla S, Hajibabaei M. 2012 DNA barcodes for everyday life: routine authentication of Natural Health Products. *Food Res. Int.* **49**, 446–452. (doi:10.1016/j.foodres.2012.07.048)
  38. Stoeckle MY, Gamble CC, Kirpekar R, Young G, Ahmed S, Little DP. 2011 Commercial teas highlight plant DNA barcode identification successes and obstacles. *Sci. Rep.* **1**, 42. (doi:10.1038/srep00042)
  39. Han J, Pang X, Liao B, Yao H, Song J, Chen S. 2016 An authenticity survey of herbal medicines from markets in China using DNA barcoding. *Sci. Rep.* **6**, 18723. (doi:10.1038/srep18723)
  40. Xin T *et al.* 2015 Survey of commercial *Rhodiola* products revealed species diversity and potential safety issues. *Sci. Rep.* **5**, 8337. (doi:10.1038/srep08337)
  41. Palhares RM, Drummond MG, Alves figueiredo Brasil BDS, Cosenza GP, Lins Brandao MDG, Oliveira G. 2015 Medicinal plants recommended by the World Health Organization: DNA barcode identification associated with chemical analyses guarantees their quality. *PLoS ONE* **10**, e0127866. (doi:10.1371/journal.pone.0127866)
  42. Little DP. 2014 Authentication of *Ginkgo biloba* herbal dietary supplements using DNA barcoding. *Genome* **57**, 513–516. (doi:10.1139/gen-2014-0130)
  43. Kumar JUS *et al.* 2015 DNA barcoding to assess species adulteration in raw drug trade of 'Bala' (genus: *Sida* L.) herbal products in South India. *Biochem. Syst. Ecol.* **61**, 501–509. (doi:10.1016/j.bse.2015.07.024)
  44. Pei N, Lian J-Y, Erickson DL, Swenson NG, Kress WJ, Ye W-H, Ge X-J. 2011 Exploring tree-habitat associations in a Chinese subtropical forest plot using a molecular phylogeny generated from DNA barcode loci. *PLoS ONE* **6**, e21273. (doi:10.1371/journal.pone.0021273)
  45. Shapcott A, Forster PI, Guymer GP, McDonald WJF, Faith DP, Erickson D, Kress WJ. 2015 Mapping biodiversity and setting conservation priorities for SE Queensland's rainforests using DNA barcoding. *PLoS ONE* **10**, e0122164. (doi:10.1371/journal.pone.0122164)
  46. Saslis-Lagoudakis CH *et al.* 2012 Phylogenies reveal predictive power of traditional medicine in bioprospecting. *Proc. Natl Acad. Sci. USA* **109**, 15 835–15 840. (doi:10.1073/pnas.1202242109)
  47. Fazekas AJ, Kesanakurti PR, Burgess KS, Percy DM, Graham SW, Barrett SCH, Newmaster SG, Hajibabaei M, Husband BC. 2009 Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Mol. Ecol. Resour.* **9**, 130–139. (doi:10.1111/j.1755-0998.2009.02652.x)
  48. Pei N *et al.* 2015 Closely-related taxa influence woody species discrimination via DNA barcoding: evidence from global forest dynamics plots. *Sci. Rep.* **5**, 15127. (doi:10.1038/srep15127)
  49. Petit RJ, Excoffier L. 2009 Gene flow and species delimitation. *Trends Ecol. Evol.* **24**, 386–393. (doi:10.1016/j.tree.2009.02.011)
  50. Naciri Y, Caetano S, Salamin N. 2012 Plant DNA barcodes and the influence of gene flow. *Mol. Ecol. Resour.* **12**, 575–580. (doi:10.1111/j.1755-0998.2012.03130.x)
  51. Percy DM *et al.* 2014 Understanding the spectacular failure of DNA barcoding in willows (*Salix*): does this result from a trans-specific selective sweep? *Mol. Ecol.* **23**, 4737–4756. (doi:10.1111/mec.12837)
  52. Twyford AD. 2014 Testing evolutionary hypotheses for DNA barcoding failure in willows. *Mol. Ecol.* **23**, 4674–4676. (doi:10.1111/mec.12892)
  53. Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S. 2015 Plant DNA barcoding: from gene to genome. *Biol. Rev.* **90**, 157–166. (doi:10.1111/brv.12104)
  54. Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S. 2015 *ycf1*, the most promising plastid DNA barcode of land plants. *Sci. Rep.* **5**, 8348. (doi:10.1038/srep08348)
  55. Hollingsworth ML *et al.* 2009 Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Mol. Ecol. Resour.* **9**, 439–457. (doi:10.1111/j.1755-0998.2008.02439.x)
  56. Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SCH. 2008 Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE* **3**, e2802. (doi:10.1371/journal.pone.0002802)
  57. Pillon Y, Johansen J, Sakishima T, Chamala S, Barbazuk WB, Roalson EH, Price DK, Stacy EA. 2013 Potential use of low-copy nuclear genes in DNA barcoding: a comparison with plastid genes in two Hawaiian plant radiations. *BMC Evol. Biol.* **13**, 35. (doi:10.1186/1471-2148-13-35)
  58. Ruhsam M *et al.* 2015 Does complete plastid genome sequencing improve species discrimination and phylogenetic resolution in *Araucaria*? *Mol. Ecol. Resour.* **15**, 1067–1078. (doi:10.1111/1755-0998.12375)
  59. Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JMM, Cronk Q. 2012 Ultra-barcoding in Cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *Am. J. Bot.* **99**, 320–329. (doi:10.3732/ajb.1100570)
  60. Nock CJ, Waters DLE, Edwards MA, Bowen SG, Rice N, Cordeiro GM, Henry RJ. 2011 Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnol. J.* **9**, 328–333. (doi:10.1111/j.1467-7652.2010.00558.x)
  61. Dong W, Xu C, Cheng T, Lin K, Zhou S. 2013 Sequencing angiosperm plastid genomes made easy: a complete set of universal primers and a case study on the phylogeny of Saxifragales. *Genome Biol. Evol.* **5**, 989–997. (doi:10.1093/gbe/evt063)
  62. Yang JB, Li DZ, Li HT. 2014 Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. *Mol. Ecol. Resour.* **14**, 1024–1031.
  63. Shi C, Hu N, Huang H, Gao J, Zhao Y-J, Gao L-Z. 2012 An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. *PLoS ONE* **7**, e31468. (doi:10.1371/journal.pone.0031468)
  64. Stull GW *et al.* 2013 A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Appl. Plant Sci.* **1**, 1200497. (doi:10.3732/apps.1200497)
  65. Straub SC, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. 2012 Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *Am. J. Bot.* **99**, 349–364. (doi:10.3732/ajb.1100335)
  66. Coissac E, Hollingsworth PM, Lavergne S, Taberlet P. 2016 From barcodes to genomes: extending the concept of DNA barcoding. *Mol. Ecol.* **25**, 1423–1428. (doi:10.1111/mec.13549)
  67. Coissac E. 2016 ORG.asm 0.2.04: A de novo assembler dedicated to organelle genome assembling. See <https://pypipython.org/pypi/ORGasm/>.
  68. Wyman SK, Jansen RK, Boore JL. 2004 Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20**, 3252–3255. (doi:10.1093/bioinformatics/bth352)
  69. Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. 2014 From algae to angiosperms – inferring the phylogeny of green plants



- (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biol.* **14**, 1–27. (doi:10.1186/1471-2148-14-23)
70. Wicke S, Schneeweiss G, dePamphilis C, Müller K, Quandt D. 2011 The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol. Biol.* **76**, 273–297. (doi:10.1007/s11103-011-9762-4)
71. Rieseberg LH, Soltis DE. 1991 Phylogenetic consequences of cytoplasmic gene flow in plants. *Evol. Trends Plants* **5**, 65–84.
72. Nichols R. 2001 Gene trees and species trees are not the same. *Trends Ecol. Evol.* **16**, 358–364. (doi:10.1016/S0169-5347(01)02203-0)
73. Nicholls JA, Pennington RT, Koenen EJM, Hughes CE, Hearn J, Bunnefeld L, Dexter KG, Stone GN, Kidner CA. 2015 Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Front. Plant Sci.* **6**, 710. (doi:10.3389/fpls.2015.00710)
74. de Sousa F, Bertrand YJK, Nylinder S, Oxelman B, Eriksson JS, Pfeil BE. 2014 Phylogenetic properties of 50 nuclear loci in *Medicago* (Leguminosae) generated using multiplexed sequence capture and next-generation sequencing. *PLoS ONE* **9**, e109704. (doi:10.1371/journal.pone.0109704)
75. Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD. 2002 Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell* **14**, 1457–1467. (doi:10.1105/tpc.010479)
76. Zhang N, Zeng L, Shan H, Ma H. 2012 Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. *New Phytol.* **195**, 923–937. (doi:10.1111/j.1469-8137.2012.04212.x)
77. Matasci N *et al.* 2014 Data access for the 1,000 plants (1KP) project. *GigaScience* **3**, 1–10. (doi:10.1186/2047-217X-3-17)
78. Mandel JR, Dikow RB, Funk VA, Masalia RR, Staton SE, Kozik A, Michelmore RW, Rieseberg LH, Burke JM. 2014 A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *Appl. Plant Sci.* **2**, 1300085. (doi:10.3732/apps.1300085)
79. Bakker FT *et al.* 2016 Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biol. J. Linnean Soc.* **117**, 33–43. (doi:10.1111/bij.12642)
80. Jiao X *et al.* 2013 A benchmark study on error assessment and quality control of CCS reads derived from the PacBio RS. *J. Data Mining Genom. Proteom.* **4**, 16008. (doi:10.4172/2153-0602.1000136)
81. Mikheyev AS, Tin MM. 2014 A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* **14**, 1097–1102. (doi:10.1111/1755-0998.12324)
82. English AC *et al.* 2012 Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768. (doi:10.1371/journal.pone.0047768)
83. Ferrarini M *et al.* 2013 An evaluation of the PacBio RS platform for sequencing and *de novo* assembly of a chloroplast genome. *BMC Genom.* **14**, 670. (doi:10.1186/1471-2164-14-670)
84. Quick J *et al.* 2016 Real-time, portable genome sequencing for Ebola surveillance. *Nature* **350**, 228–232. (doi:10.1038/nature16996)
85. 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 Res.* **41**, e129. (doi:10.1093/nar/gkt371)
86. Bock DG, Kane NC, Ebert DP, Rieseberg LH. 2014 Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. *New Phytol.* **201**, 1021–1030. (doi:10.1111/nph.12560)
87. Malé PJG *et al.* 2014 Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Mol. Ecol. Resour.* **14**, 966–975.
88. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379. (doi:10.1371/journal.pone.0019379)
89. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. 2007 Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* **17**, 240–248. (doi:10.1101/gr.5681207)
90. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008 Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376. (doi:10.1371/journal.pone.0003376)
91. Rubin BER, Ree RH, Moreau CS. 2012 Inferring phylogenies from RAD sequence data. *PLoS ONE* **7**, e33394. (doi:10.1371/journal.pone.0033394)
92. Cruaud A *et al.* 2014 Empirical assessment of RAD sequencing for interspecific phylogeny. *Mol. Biol. Evol.* **31**, 1272–1274. (doi:10.1093/molbev/msu063)
93. Cariou M, Duret L, Charlat S. 2013 Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecol. Evol.* **3**, 846–852. (doi:10.1002/ece3.512)
94. Ekblom R, Galindo J. 2010 Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**, 1–15. (doi:10.1038/hdy.2010.152)
95. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015 The arabidopsis information resource: making and mining the 'gold standard' annotated reference plant genome. *Genesis* **53**, 474–485. (doi:10.1002/dvg.22877)
96. Li R *et al.* 2010 The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317. (doi:10.1038/nature08696)
97. VanBuren R *et al.* 2015 Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527**, 508–511. (doi:10.1038/nature15714)
98. Kelly LJ, Leitch AR, Fay MF, Renny-Byfield S, Pellicer J, Macas J, Leitch IJ. 2012 Why size really matters when sequencing plant genomes. *Plant Ecol. Divers.* **5**, 415–425. (doi:10.1080/17550874.2012.716868)
99. Kelly LJ *et al.* 2015 Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytol.* **208**, 596–607. (doi:10.1111/nph.13471)