Letter

Genome skimming for next-generation biodiversity analysis

Steven Dodsworth[1,2]

[1] School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK
[2] Department of Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond TW9 3DS, UK

High-throughput sequencing technologies have revolutionised the ease with which genomic data can be obtained for any plant species, from trees to bryophytes, regardless of phylogenetic proximity to model species or even the ease of cultivation. This is creating an impact on all types of biodiversity study, from phylogenomic and systematic studies through to population genetic, barcoding, and ecological investigations. A plethora of different approaches are available that differ in the expertise and time required for both the preparation of genomic DNA (gDNA) for sequencing and subsequent bioinformatic analysis of read data (Figure 1). At the deep end of approaches are hybridisation methods using developed (known) bait sequences, transcriptome sequencing, and reduced representation methods, such as RAD-seq. By contrast, genome skimming is by far one of the simplest methodologies, involving random sampling of a small percentage of total gDNA. This approach has been used successfully at varying taxonomic levels, for intraspecific 'ultra-barcoding', intergeneric and family-wide phylogenomic analyses [1–4].

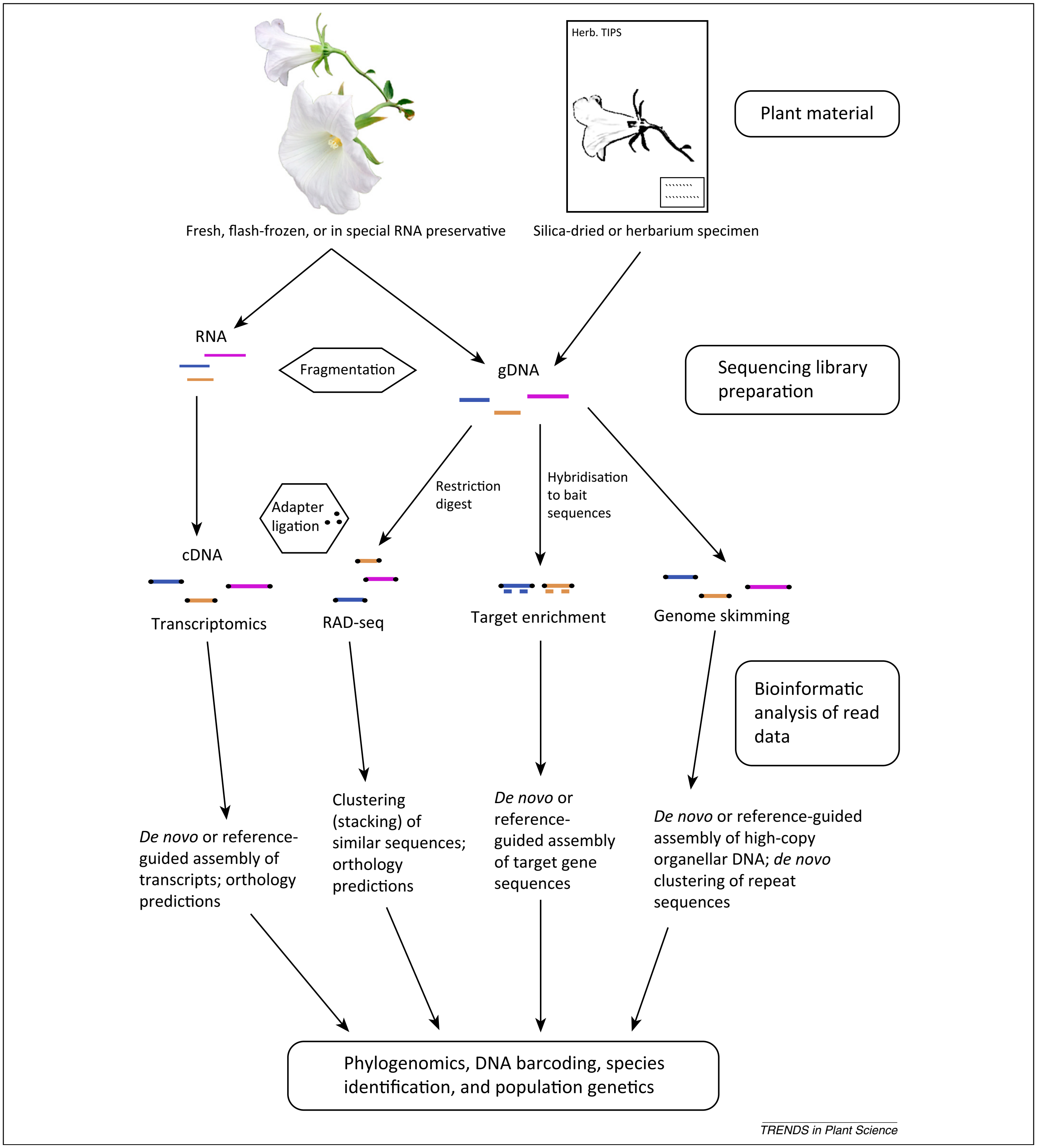
The term 'genome skimming' was first coined by Straub *et al*. [5] as a way of 'navigating the tip of the genomic iceberg'; that is, shallow sequencing of gDNA that results in comparatively deep sequencing of the high-copy fraction of the genome (plastome, mitogenome, and repetitive elements). This is a notable expansion of traditional phylogenetic markers used in plants, particularly the nuclear ribosomal internal transcribed spacer (nrITS) and various plastid markers (e.g., *rbcL* and *matK* genes: the plant DNA barcodes). Whole-plastid genomes extend this to approximately 100 genes, are useful markers for phylogenetic inference, and, therefore, are often the single goal of high-throughput systematics studies [2,6,7]. However, the amount of plastid DNA present in a particular gDNA sample depends on the tissue sampled, developmental stage, and species-specific factors. One important consideration in plants is genome size: as genome size increases, the proportion of organellar DNA in a sample will concomitantly decrease. As a result, plastid DNA can vary substantially between species and/or samples, and even within species from 0.4% to 29.5% of total extracted DNA [5,8]. Healthy mature (but not senescing) leaves are likely to hit the optimum amount of plastid DNA per cell, because very young leaves contain less plastid DNA [8] and, as leaves senesce, plastid DNA typically declines in abundance [9].

This potential caveat of a variable amount of plastid sequence reads is somewhat negated by the other data present, particularly nuclear repeats. The full ribosomal cistron, containing the much-used nrITS sequences, is well known to plant systematists. However, there are many other nuclear repeats in the data, and these will always be present in genome-skimming data sets, probably without as much variability in abundance that is a problem with plastome DNA. Nuclear repeats may provide other valuable phylogenetic information that can now be easily accessed using fast clustering approaches [10]. Another advantage of genome skimming is that it may be particularly useful for degraded gDNA. A vast, as yet relatively untapped, resource of genomic data is held in museum collections and herbaria. Unfortunately, gDNA extracted from herbarium and museum specimens is often highly degraded, owing to a combination of age and original preservation methods [11,12]. High-copy regions of the genome (including organellar genomes) are still present in such samples and, as such, methods focussing on sequencing these regions are likely to be more successful than those focussing on low-copy regions of the genome. However, there are likely to be skews in the data due to degradation that may require novel analytical methods.

A crucial aim in molecular ecology and biodiversity studies is to sequence 'live' in the field. One of the latest advents in high-throughput sequencing technologies, nanopore sequencing, is set to promise just this. Current developments in nanopore sequencing by Oxford Nanopore Technologies (ONT) have made this possibility of sequencing in the field a reality [13]. Recently, a team of Italian scientists discovered a potentially new species of frog by using the MinION palm-sized sequencer in the rainforests of Tanzania [14]. Their analysis and other initial users of this technology have tended to rely on first amplifying regions by conventional PCR. However, this requires a less-than-portable PCR machine along with further reagents for sample preparation. A fundamental goal is to relieve the burden of sample preparation altogether and simply drop gDNA into the device, something that ONT are promising in the near-future with their Voltrax device, which processes samples automatically and docks on top of the portable sequencer. Surely genome skimming, sequencing the high-copy portion of nuclear DNA and orga-

Corresponding author: Dodsworth, S. (steven.dodsworth@qmul.ac.uk).
Keywords: genome skimming; phylogenomics; phylogenetics; systematics; high-throughput sequencing; DNA barcoding.

1360-1385/
http://dx.doi.org/10.1016/j.tplants.2015.06.012


Herb. TIPS
Plant material
Fresh, flash-frozen, or in special RNA preservative
Silica-dried or herbarium specimen
RNA
Fragmentation
gDNA
Sequencing library preparation
Restriction digest
Hybridisation to bait sequences
Adapter ligation
cDNA
Transcriptomics
RAD-seq
Target enrichment
Genome skimming
Bioinformatic analysis of read data
De novo or reference-guided assembly of transcripts; orthology predictions
Clustering (stacking) of similar sequences; orthology predictions
De novo or reference-guided assembly of target gene sequences
De novo or reference-guided assembly of high-copy organellar DNA; *de novo* clustering of repeat sequences
Phylogenomics, DNA barcoding, species identification, and population genetics

TRENDS in Plant Science

Figure 1. Summary of current high-throughput sequencing methods applied to evolutionary and ecological studies. Abbreviations: gDNA, genomic DNA; cDNA, complementary DNA.

nellar DNA, and running the small portable device for a relatively short length of time, would be a near-perfect solution for next-generation biodiversity analysis?

References

1 Kane, N. *et al.* (2012) Ultra-barcoding in cacao (Theobroma spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *Am. J. Bot.* 99, 320–329

2 McPherson, H. *et al.* (2013) Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. *BMC Ecol.* 13, 8

3 Besnard, G. *et al.* (2013) Phylogenomics and taxonomy of Lecomtelleae (Poaceae), an isolated panicoid lineage from Madagascar. *Ann. Bot.* 112, 1057–1066

4 Malé, P.J.G. *et al.* (2014) Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Mol. Ecol. Resour.* 14, 966–975

5 Straub, S.C.K. *et al.* (2012) Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *Am. J. Bot.* 99, 349–364

6 Van der Merwe, M. *et al.* (2014) Next-Gen phylogeography of rainforest trees: Exploring landscape-level cpDNA variation from whole-genome sequencing. *Mol. Ecol. Resour.* 14, 199–208

7 Li, X. *et al.* (2014) Plant DNA barcoding: from gene to genome. *Biol. Rev.* 90, 157–166

8 Rauwolf, U. *et al.* (2010) Variable amounts of DNA related to the size of chloroplasts III. Biochemical determinations of DNA amounts per organelle. *Mol. Genet. Genomics* 283, 35–47

9 Rowan, B.A. and Bendich, A.J. (2009) The loss of DNA from chloroplasts as leaves mature: fact or artefact? *J. Exp. Bot.* 60, 3005–3010

10 Dodsworth, S. *et al.* (2015) Genomic repeat abundances contain phylogenetic signal. *Syst. Biol.* 64, 112–126

11 Staats, M. *et al.* (2013) Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLoS ONE* 8, e69189

12 Särkinen, T. *et al.* (2012) How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLoS ONE* 7, 1–9

13 Laver, T. *et al.* (2015) Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* 3, 1–8

14 Hayden, E.C. (2015) Pint-sized DNA sequencer impresses first users. *Nature* 521, 15–16