Forum Article

Hyb-Seq for flowering plant systematics

Steven Dodsworth[1,2†*], Lisa Pokorny[1†], Matthew G. Johnson[3,4†], Jan T. Kim[1], Olivier Maurin[1], Norman J. Wickett[4,5], Felix Forest[1], William J. Baker[1]

[1]Royal Botanic Gardens, Kew, Richmond TW9 3AE, Surrey, UK.
[2]School of Life Sciences, University of Bedfordshire, University Square, Luton LU1 3JU, UK.
[3]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA
[4]Chicago Botanic Garden, Glencoe, IL 60022, USA
[5]Program in Plant Biology and Conservation, Northwestern University, Evanston, IL 60208, USA

[†]Authors contributed equally
[*]Correspondence: steven.dodsworth@beds.ac.uk

Keywords

High-throughput sequencing – molecular systematics – phylogenetics – Hyb-Seq – sequence capture – angiosperms – tree of life – genomics

Abstract

High-throughput DNA sequencing (HTS) presents great opportunities for plant systematics, yet genomic complexity needs to be reduced for HTS to be effectively applied. We highlight Hyb-Seq as a promising approach, especially in light of the recent development of probes enriching 353 low-copy nuclear genes from any flowering plant taxon.

High-throughput sequencing approaches and plant systematics

Current developments in DNA sequencing, collectively termed high-throughput sequencing (HTS) technologies, permit many orders of magnitude more DNA data to be routinely collected compared to standard Sanger sequencing. This has made whole genome sequencing of diverse plant taxa much more accessible, including both flowering and non-flowering land plant lineages. However, challenges prevail: plant genome size varies enormously [1], genome assembly is often non-trivial for even the smallest plant genomes, and the cost per high-quality genome sequence is still significant. This means that, at least for the time being, methods are needed to reduce genomic complexity. This is especially the case for phylogenetics and systematics, in order to find an optimal amount of sequencing effort per sample whilst reaping the benefits of increased data. In this article, we propose Hyb-Seq as one of the most promising approaches for plant systematists currently, and particularly in light of a recent set of probes that target low-copy regions of the nuclear genome across flowering plants (angiosperms).

Systematics is primarily concerned with evolutionary relationships and natural classification, and as such producing reliable phylogenetic frameworks is often of primary concern. This is not the same as genomic studies, where detailed dissection of phenotypic traits or speciation processes may be the main goal—though there is a strong overlap between these fields. Phylogenetic data requires a constant trade-off between the depth (characters as DNA base pairs) and breadth (number of taxa) of data collected. Different evolutionary questions may demand different compromises on the depth-breadth spectrum. This is also a tension between an idealised data source (a complete nuclear genome sequence) and one that is easier and quicker to produce but far less information-rich (a small DNA barcode of a few hundred base pairs). Such examples lie at either end of a continuum of DNA sequencing tactics, making it difficult to find an optimal approach (Table 1).

Herbarium specimens are the foundation of taxonomic studies in plants. Herbarium DNA is usually highly fragmented and often contaminated, making PCR-based approaches challenging [2,3]. HTS can surmount these difficulties as all native DNA fragments present can potentially be sequenced [3,4], although different approaches have their advantages and disadvantages (see below).

Genome Skimming

Simple approaches such as genome skimming [4] remain popular, although recovery of orthologous nuclear regions for sequence alignment is limited with these techniques. Whilst organellar genomes (particularly plastid genomes) are easily reconstructed from such data, their histories reflect patterns associated with matrilineal genealogy/geography or other aspects of organelle biology. As such phylogenetic inference based on plastid or organellar data may not necessarily reflect the evolutionary history of the taxa in question (for a comprehensive view of plastid evolution, see [5]). Ribosomal DNA is easily recovered, although not always highly variable and concerted evolution can produce incongruent topologies. Other repetitive elements (e.g. satellite DNA, transposable elements) can be easily quantified from a genome skim, but sequence divergence of such repeats is low. Repeat abundance and repeat sequence similarity can be used instead of sequence alignment for phylogenetic reconstruction [6] although these are very different approaches, both conceptually and practically.

RAD-Seq

Restriction site-associated DNA sequencing (RAD-Seq or similar Genotyping-by-Sequencing approaches; GBS) is a method to sequence DNA next to restriction sites. The loci are essentially random, although partially selection for particular genomic contexts (e.g. genic regions) is possible using methylation-sensitive enzymes [7]. RAD-Seq holds particular promise at shallow scales, for resolving recent radiations and population-level sampling [8], where a large number of single nucleotide polymorphisms (SNPs) help. RAD-Seq loci are often short, however, and not always easy to annotate without a high-quality reference genome. As genomic DNA is cut with enzymes, high molecular weight DNA is required. Recent silica-dried collections therefore work well as do very recent herbarium specimens but degraded DNA from older herbarium specimens will not work. Due to the variability of restriction sites between taxa, particularly over larger evolutionary distances, securing enough homologous loci is difficult at deeper (or variable) phylogenetic scales. This also means that RAD-Seq data in public repositories may not be a very usable resource (e.g. as a source of outgroup sequences from related taxa).

RNA-Seq

Transcriptomics requires high-quality RNA from samples, which usually means flash-frozen using liquid nitrogen or dry ice or using pricey preservative liquids designed to preserve RNA in the field and requiring -80 °C storage. Resulting data will include all expressed genes in that particular sample, which makes RNA-Seq ideal for obtaining large numbers of protein-coding genes. Due to differences in expression throughout the plant, though, a variety of tissues should ideally be used (e.g. flower, root, leaf). There are some obvious caveats to this approach: (i) it requires healthy living plant tissue and access to preservatives/freezers; and (ii) it may require a variety of tissues; and (iii) it remains relatively expensive per sample (Table 1).

Sequence capture, target enrichment and Hyb-Seq approaches

Bait design

Sequence capture approaches are becoming increasingly popular as a method of reducing genomic complexity, exploiting “baits” (probes) to enrich specific target regions (loci) from total DNA. This approach has been variously referred to as bait hybridisation, target enrichment, sequence/target/hybrid capture, Hyb-Seq, or other combinations of such terms. A common feature is the use of pre-designed RNA or DNA bait sequences, developed from pre-existing genomic information, such as a closely-related genome sequence or transcriptome data. Target loci are often nuclear protein-coding sequences or other conserved genomic regions, such as ultra-conserved elements (UCEs—in animals and fungi). Typically, low-copy (ideally single-copy) genes are chosen for phylogenetic purposes, thus minimising any orthology issues later on. In many cases, however, multigene families are also included [e.g. 9], particularly where those genes have known functions of biological interest to the groups being studied (e.g. photosynthetic transitions, or transcription factors involved in morphological diversity).

If protein-coding regions are targeted, phylogenetic inference can employ explicit models that account for different rates of evolution based on codon position. Such explicit positional information is often required for reliable inference at deeper phylogenetic scales [10]. Codon positions are often difficult to infer using RAD-Seq data, protein-coding nuclear data are lacking in genome skims, and RNA-Seq is expensive. Hyb-Seq can provide protein-coding data at a fraction of the cost, and a compromise point where these other approaches fall down.

Generalised workflow

Genomic DNA extracts are first turned into libraries of genomic fragments. The RNA/DNA baits are subsequently hybridized to target loci in genomic libraries. Bait-bound DNA is then separated from the mixture, e.g. by using streptavidin-coated magnetic beads that bind biotinylated baits (and bait-bound DNA), that can then be separated simply with a magnet (Figure 1). DNA fragments not bound to baits are discarded through a series of washing steps, and the result is a pool of fragments enriched for particular target sequences (Figure 1).

Effective recovery of target loci can be achieved even with surprisingly low levels of enrichment, as low as 10% of the sequence reads [9]. Consequently, there can be abundant off-target reads that include high-copy DNA regions, such as repetitive DNA, the ribosomal operon, and organellar DNA from plastids and mitochondria (Figure 1). This off-target fraction is similar to a genome skim [4], or low-coverage whole-genome sequencing, and can also be exploited for systematic analyses [11]. Moreover, regions adjacent to the target loci (known as the "splash zone") are also recovered (Figure 1), often including intronic regions, which may be highly variable and therefore valuable at shallower phylogenetic levels [12,13].

Hyb-Seq

The term Hyb-Seq was initially proposed by Weitemier et al. (2014; [12]) to consider the explicit use of both the on-target reads (i.e. enriched gene sequences) and the off-target fraction. In recent years, the term Hyb-Seq has had slightly different meanings, such as mixing the enriched and unenriched (native) libraries [11], or explicitly sequencing both enriched and unenriched sets of libraries separately. The fundamental meaning remains the same—utilisation of both low-copy enriched nuclear sequences and high-copy unenriched ones such as plastid and ribosomal DNA.

The unenriched category notably and conveniently includes markers that have been traditionally used for decades in plant systematics, the currently used plant DNA barcodes—*rbcL*, *matK*, *trnH-psbA* spacer (plastid genome) and nrITS of ribosomal DNA. Sequencing these loci will facilitate the ongoing global synthesis of plant systematic data for a variety of use cases. Hyb-Seq has been successfully used in a number of groups at varying levels of phylogenetic depth [e.g. 11,12]; it has also been used very effectively with herbarium samples, including those over 100 years old and spanning the diversity of angiosperms [11,14].

Enriching a core set of genes in flowering plants and future potential

Angiosperms-353 bait set

Probes for sequence capture have traditionally been designed for specific plant groups of interest. The design of such a kit requires access to (or production of) genomic resources and at least some bioinformatic expertise. Recent publication of an angiosperm-wide set of baits makes Hyb-Seq a great deal more accessible for flowering plants and alleviates part of the financial and bioinformatic burden [4]. Johnson et al. (2018; [15]) have developed a probe set that targets 353 low-copy orthologous nuclear genes in angiosperms, derived from an alignment of low-copy genes across all green plants by the 1000 Plant Transcriptomes Initiative or OneKP project (onekp.com). Their approach includes the use of up to 15 variants for each of the 353 gene loci (approx. 230 Kbp of nuclear sequence), in order to capture sequence diversity across angiosperms with one single kit (Angiosperms-353, available at arborbiosci.com/products/mybaits-plant-angiosperms, catalog #3081XX). Including variants means that, on average, DNA from 95% of angiosperm species should hybridise to one or more gene variants with ≤ 30% divergence between the sample and the target sequence. Importantly, hybridisation is reported to be efficient below such a threshold.

Future potential

This means that this kit should work for any of the 300,000 currently estimated angiosperm species, distributed in 416 families, and which dominate terrestrial ecosystems globally. Johnson et al. [15] show very promising data for 42 samples taken from across the angiosperms, with no obvious systematic/taxonomic biases, and potential phylogenetic signal at various levels.

The Angiosperms-353 kit has enormous potential for studies that combine deep and shallow-level systematic studies. There is also promise as a powerful new tool in the fields of molecular and community ecology (e.g. discovering the types of pollen carried by pollinators, community assembly, or characterising habitats through molecular sampling). This is potentially possible by building a database of a common set of hundreds of genes per sample. Such a set of core genes may even be a nuclear solution for the “next generation” flowering-plant DNA barcode.

References

1. Pellicer, J. *et al*. (2018) Genome size diversity and its impact on the evolution of land plants. *Genes* 9, 88.

2. Särkinen, T. *et al*. (2012) How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLoS One*, e43808.

3. Bakker, F.T. (2017) Herbarium genomics: skimming and plastomics from archival specimens. *Webbia* 72, 35-45.

4. Dodsworth, S. (2015) Genome skimming for next-generation biodiversity assessment. *Trends in Plant Science* 20, 525-527.

5. Gitzendanner, M.A. *et al*. (2018) Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *Am. J. Bot.*

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

7. Elshire, R.J. *et al*. (2011) A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. *PLoS One* 6, e19379.

8. Paun, O. *et al*. (2015) Processes driving the adaptive radiation of a tropical tree (*Diospyros*, Ebenaceae) in New Caledonia, a biodiversity hotspot. *Systematic Biology* 65, 212-227.

9. Moore, A.J. *et al*. (2017) Targeted enrichment of large gene families for phylogenetic inference: Phylogeny and molecular evolution of photosynthesis genes in the Portullugo clade (Caryophyllales). *Syst. Biol.* 67, 367-383.

10. Wickett, N.J. *et al*. (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. *PNAS* 111, E4859-68.

11. Villaverde, T. *et al*. (2018) Bridging the micro and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. *New Phytologist* 220, 636-650.

12. Weitemier, K. *et al*. (2014) Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Appl. Plant Sci.* 2

13. Johnson, M.G. *et al*. (2016) HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.* 4, 1600016-1600018

14. Hart, M.L. *et al*. (2016) Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon* 65, 1081-1092.

15. Johnson, M.G. *et al*. (2018) A universal probe set for sequence capture of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology* 68, 594-606.

Acknowledgements

This research was supported by grants from the Calleva Foundation, the Sackler Trust and the Garfield Weston Foundation to the Royal Botanic Gardens, Kew.

Table 1. Comparison of high-throughput sequencing approaches for plant systematics: advantages and disadvantages[a]

Phylogenomics approach	Genomic resources required	Initial bioinformatic investment	Ultimate bioinformatic investment	Initial laboratory cost	Ultimate cost per sample	Low-copy nuclear genes retrieved
Genome skimming	No	None	Medium	Low	Medium	No/Limited
RAD-Seq	No, but helpful	Medium	High	High	Low	No/SNPs
RNA-Seq	No, but helpful	Low	High	Low	High	Yes-thousands
Hyb-Seq	Varies[b]	High[b]	Medium	Low[b]	Medium	Yes-variable

[a]Initial costs include the one-time or limited purchase of expensive consumables (e.g. biotinylated baits or adapter sequences). Boxes are highlighted from unfavourable (red) to favourable (green) under each column.

[b]If designing new kit(s) genome or transcriptome resources are required, otherwise readily available kits exist for different groups of plants as well as angiosperms as a whole (Angiosperms-353) and are much cheaper than designing a new custom bait set.

Figure 1. Simplified schematic representing the main steps in a typical Hyb-Seq workflow: (i) Libraries of double-stranded DNA fragments are prepared from genomic DNA; (ii) Libraries are denatured (single-stranded) and bound to biotinylated probes/baits; (iii) streptavidin-coated magnetic beads bind to the biotinylated bait-DNA hybrids, these are bound to a magnet, and other DNA fragments are washed off; (iv) baited-DNA is PCR-ed and removed from the beads for sequencing. Target DNA sequences are in dark blue and non-target sequences are in orange. Hyb-Seq has the potential to recover both "splash zone" sequences close to targets (edges of dark blue sequences in orange, e.g. introns) as well as some completely off-target sequences (orange blocks, e.g. plastid DNA), as indicated in the final sequencing library (iv).