Plant phylogenomics based on genome-partitioning strategies: Progress and prospects

Xiangqin Yu a, Dan Yang a, c, Cen Guo b, c, Lianming Gao a, *

a Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan, 650201, China
b Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan, 650201, China
c Kunming College of Life Science, University of Chinese Academy of Sciences, Kunming, Yunnan, 650201, China

A R T I C L E I N F O

Article history:
Received 24 April 2018
Received in revised form 26 June 2018
Accepted 27 June 2018
Available online 30 June 2018
(Editor: Zhekun Zhou)

Keywords:
Plant phylogenomics
Next-generation sequencing
Whole-genome sequencing
Genome skimming
RAD-Seq
Targeted capture

A B S T R A C T

The rapid expansion of next-generation sequencing (NGS) has generated a powerful array of approaches to address fundamental questions in biology. Several genome-partitioning strategies to sequence selected subsets of the genome have emerged in the fields of phylogenomics and evolutionary genomics. In this review, we summarize the applications, advantages and limitations of four NGS-based genome-partitioning approaches in plant phylogenomics: genome skimming, transcriptome sequencing (RNA-seq), restriction site associated DNA sequencing (RAD-Seq), and targeted capture (Hyb-seq). Of these four genome-partitioning approaches, targeted capture (especially Hyb-seq) shows the greatest promise for plant phylogenetics over the next few years. This review will aid researchers in their selection of appropriate genome-partitioning approaches to address questions of evolutionary scale, where we anticipate continued development and expansion of whole-genome sequencing strategies in the fields of plant phylogenomics and evolutionary biology research.

Copyright © 2018 Kunming Institute of Botany, Chinese Academy of Sciences. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Phylogenomics is a new and exciting synthesized discipline (Delsuc et al., 2005; Eisen, 1998) that is at the intersection of evolution and genomics (Eisen and Fraser, 2003). The main goal of phylogenomics is to infer species relationships using genomic data, as well as to gain knowledge of the mechanisms of molecular evolution based on the evolutionary history of species (Philippe et al., 2005). Understanding phylogenetic relationships between organisms is a prerequisite of almost all evolutionary studies (Delsuc et al., 2005; Zhang et al., 2012), although many plant phylogenies based on traditional DNA-fragments remain unresolved at all evolutionary scales due to a lack of informative sites. Phylogenomics can potentially resolve species relationships by making use of vast sequence data as well as gene order, insertions and deletions (indels), retroposon integrations, and gene fusion and fission events (Rokas and Holland, 2000). Access to genomic data could also potentially alleviate previous problems of phylogenetics that resulted from stochastic error (limitation of sampling few genes) by expanding the number of characters (Delsuc et al., 2005). Thus, phylogenomics provides a window for better understanding the evolutionary relationships of plants using genome-scale data and generating a more robust picture of the next generation Tree of Life.

* Corresponding author. Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, 132# Lanhei Road, Kunming, 650201, Yunnan, China. Fax: +86 0871 65225286.
E-mail address: gaolm@mail.kib.ac.cn (L. Gao).
Peer review under responsibility of Editorial Office of Plant Diversity.
From first generation (Sanger) DNA sequencing to the second (massively parallel) and third (real-time, single-molecule) generation DNA sequencing, gathering genomic data has become much more convenient and cost-effective (Shendure et al., 2017). Among the three genomes of plants, the plastid genome evolves more rapidly and has lower inter- and intramolecular recombination rates than the mitochondrial genome (Lonsdale et al., 1988; Palmer and Herbon, 1988). Furthermore, the plastid genome can be more easily sequenced than the nuclear genome. Despite earlier debate on genome-based phylogenies, Martin et al. (2005) argued for the critical role that the plastid genome plays in plant phylogenetics. Sequencing the plastid genome was the most common approach in the early stages of plant phylogenetic and evolution research, providing an efficient method for investigating the evolutionary relationships and basal lineages of angiosperms (Goremykin et al., 2003, 2004; Jansen et al., 2007; Moore et al., 2007).

Since 2015, improved NGS technologies, lower genome sampling costs, and the development of statistical methods have greatly expanded the use of genomic data in plant phylogenetics (Barrett et al., 2016; Gao et al., 2010). Various genome-partitioning strategies to sequence selected subsets of the genome, such as genome skimming, transcriptome sequencing (RNA-seq), restriction site associated DNA sequencing (RAD-Seq), and targeted capture have emerged as powerful alternatives to whole-genome sequencing (WGS) in ecological and evolutionary genomic studies (Jones and Good, 2016) as well as studies in plant phylogenetics. For example, as of April 2018, more than 2800 records of whole plastid genomes of plants have been deposited in GenBank. In most cases, sequences were obtained by long-range polymerase chain reaction (PCR) (Yang et al., 2014) and genome skimming (Straub et al., 2012) with NGS. Based on such data, plastid phylogenomics has undoubtedly been serving as an effective approach for uncovering deep-level relationships of intractable and even rapidly radiating plant groups (Barrett et al., 2013; Ma et al., 2014; Ross et al., 2015; Wysocki et al., 2015). In addition, nuclear and mitochondrial genomes are expected to play an increasing role in plant phylogenetics in the future (Liu et al., 2014; Vargas et al., 2017; Zeng et al., 2014; Zhang et al., 2012). The international multidisciplinary consortium “1000 Plants (1KP) project” aims to generate transcriptome data from over 1300 green plants, including all of the major lineages across the Viridiplantae clade, underscoring the considerable effort this team has made to investigate the value of nuclear genomic data for plant phylogenomic analyses (Matasci et al., 2014). Furthermore, making use of the whole nuclear genome or data from resequencing for plant phylogenomics and population genomics is also at the forefront of many studies (Sollars et al., 2016; Teh et al., 2017; Zhang et al., 2017a). We have now reached the point where these approaches to answering fundamental evolutionary questions have transformed our research to an unprecedented degree.

2. The application of genomic data in plant phylogenetics

Over the past few years, NGS-based genomic data have made a profound impact on phylogenetics (Jarvis et al., 2013; Misof et al., 2014). Several genome-partitioning strategies to sequence selected subsets of the genome, including genome skimming, RNA-seq, RAD-Seq, and targeted capture, have emerged as powerful tools in plant phylogenomics. Here we summarize the applications, advantages, and limitations of four NGS-based genome-partitioning approaches in plant phylogenomics (Table 1). We hope this review will help researchers choose the appropriate approaches to address phylogenetic questions at various evolutionary scales.

2.1. Phylogenomics using plastid genomes

Land plant plastid genomes share the typical quadripartite structure including two rRNA-containing inverted repeats (IRs) and two unequal single-copy regions (Raubes and Jansen, 2005). The plastid genome size of land plants ranges from 11 kb to 217 kb (Bellot and Renner, 2015; Guisinger et al., 2010). Even though some rearrangements occur in certain lineages, the structure of land plant plastid genomes is generally conserved (Gao et al., 2010). The first two plastid genomes were determined by constructing a set of overlapping restriction endonuclease fragments and Sanger sequencing (Ohyama et al., 1986; Shinozaki et al., 1986). Faster and more cost-effective approaches were then developed, including 1) shearing, cloning and sequencing the isolated pure cpDNA; 2) amplification using long PCR; and 3) construction of bacterial artificial chromosome (BAC) or fosmid libraries (Jansen et al., 2005). Using conserved primers based on available plastid genomes, long range PCR was an effective tool for obtaining whole genomes for phylogenetic analyses such as identifying basal angiosperm lineages (Goremykin et al., 2003, 2004). However, these studies mostly rely on traditional Sanger sequencing, a process that was time-consuming and expensive. With the advent of NGS in 2005, followed by the development of library-construction-based NGS platforms (such as Roche 454, Solexa, SOLID and Heli-cos), sequencing entered a new era characterized by high-throughput and cost-efficiency (Shendure et al., 2017). Angiosperm phylogenies were then greatly improved by including more genes from plastid genomes and extending taxa sampling, mainly through sequencing whole plastid genomes (Moore et al., 2007, 2010). By integrating more plastid genomes as templates, Yang et al. (2014) and Zhang et al. (2016) reported nine and fifteen novel universal primer pairs for amplification of whole plastid genomes of angiosperms, respectively. This approach was subsequently used in phylogenomic analyses of several plant lineages such as Theaceae, Rosaceae and Cornales (Fu et al., 2017; Yu et al., 2017; Zhang et al., 2017b). The disadvantage of this approach is the need of high quality genomic DNA isolated from fresh material or high quality material quickly dried in silica-gel after collection.

Genome skimming was proposed as a way of ‘navigating the tip of the genomic iceberg’ (Nock et al., 2011). This approach consists of shallow sequencing of genomic DNA that results in comparatively deep sequencing of the high-copy fraction of the genome (Straub et al., 2012). Genome skimming is an efficient approach for obtaining the complete plastid genome (ptDNA), a large fraction of the mitochondrial genome (mtDNA) and the nuclear ribosomal cluster (nrDNA). It is cost-effective, and can tolerate low-quality samples (fresh, silica-gel-preserved leaves or even Herbarium specimens up to 146 years old) (Bakker et al., 2015; Dodsworth, 2015). As such, the use of genome skimming has contributed to numerous advances in our understanding of species relationships across a broad phylogenetic range of taxa. For example, shotgun-sequencing-based genome skimming of the pantropical tree family Chrysobalanaceae yielded more robust phylogenetic relationships than previous studies (Male et al., 2014; Barrett et al., 2016) obtained 39 plastomes using genome skimming to investigate deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. The pattern of reticulate evolution in a species-rich and recently diverged Andean genus Diplastephium was revealed by integrating phylogenetic signal of genomic regions with different inheritance patterns using genome skimming and deRRAdseq (Vargas et al., 2017).

Organellar genomes are generally inherited unparentally; only 14% of angiosperms inherit plastids biparentally (Corriveau and Coleman, 1988). Accordingly, Gitzendanner et al. (2018) recently
suggested that the plastid genome provides only one perspective on plant evolutionary history; thus, a plastid-based tree should not be blindly accepted as the backbone tree for Viridiplantae. Organelle capture–introgression of the organelar genomes from one species into another–may cause phylogenetic inconsistencies between organelar and nuclear trees (Huang et al., 2014; Stegemann et al., 2012; Tisitrone et al., 2003; Yi et al., 2015). Recombination and gene conversion that have occurred in the plastid genome might also introduce biases and errors to phylogenetic reconstruction (Huang et al., 2014). Sullivan et al. (2017) even reported the extreme case of interspecific plastome recombination in *Picea* (Pinaceae), which might result in discordant plastid phylogenies. In contrast, nuclear genes are biparentally inherited and contain abundant genetic information, thus more data from nuclear genomes are needed to provide alternative evidence for phylogenetic relationship reconstructions (Lee et al., 2011; Zimmer, 2013).

### 2.2. Phylogenomics using transcriptome sequencing

Comparative genomic approaches among model species have been successfully used to identify a large set of orthologous nuclear markers that can now be used in phylogenetic studies, opening new avenues for molecular systematics (Duarte et al., 2010; Soltis et al., 2013). RNA-seq of two or three species per clade of the studied taxa produces a large number of single-copy candidate loci that can be screened for substitution rate, ease of amplification, and position of introns in other taxa (Harrison and Kidner, 2011). Alternatively, transcriptomes can directly be used for phylogenetic analyses. Zhang et al. (2012) used genome comparisons between seven angiosperms and one moss species to identify 1083 highly conserved, low-copy nuclear genes, five of which were valuable in reconstructing a highly resolved angiosperm phylogeny. While the angiosperm phylogeny that was reconstructed using these five genes was largely congruent with phylogenies previously inferred from organellar genes, several new placements were uncovered for some lineages. Integrating 26 new transcriptomes with previously reported orthologous genes, 59 carefully selected low-copy nuclear genes were used to build a highly supported deep-level (among eight clades) angiosperm phylogeny. Molecular clock estimates of mesangiospermae diversification have been used to illuminate a possible link between origins of some insects and the early angiosperm radiation (Zeng et al., 2014). Furthermore, orthologous nuclear genes derived from transcriptomes were used to uncover robust phylogenies for eudicot, Caryophyllales, and several species-rich families (Rosaceae, Brassicaceae, and Asteraceae) (Huang et al., 2015, 2016; Xiang et al., 2016; Yang et al., 2015; Zeng et al., 2017b).

The 1000 Plant Genomes Project (1KP), consisting of transcriptomes from over 1300 species representing the diversity of green plants, is the first international collaboration on a large-scale transcriptome sequencing project for plants, and has provided evidence for the resolution of phylogenetic uncertainties (Granados Mendoza et al., 2015; Matasci et al., 2014). For instance, the origin and early diversification of land plants has also been investigated through phylotranscriptomic (including 1KP data) analysis of 852 nuclear genes and 1,701,170 aligned sites (Wickett et al., 2014). However, transcriptomics requires living tissue for RNA extraction, and thus many existing tissue collections are unusable (Soltis et al., 2013). Therefore, a substantial amount of effort in future studies will be placed on resampling. In addition, RNA should be sampled from the same types of tissue and from individuals at the same life-history stage to obtain as many orthologous loci across samples as possible (Leemans and Leemans, 2013).

### 2.3. Phylogenomics using restriction site-associated DNA sequencing

Restriction site-associated DNA sequencing (RAD-Seq) and its related methods rely on the conservation of enzyme recognition sites (e.g., GBS: Genotyping-By-Sequencing; SLAF-seq: Specific-Locus Amplified Fragment Sequencing; SBG: Sequence-Based Genotyping). This technique can be used to survey hundreds or thousands of unlinked genetic markers adjacent to restriction sites from the nuclear genome (Baird et al., 2008; Elshire et al., 2011; Peterson et al., 2012). This approach mainly involves digestion of genomic DNA samples with restriction enzymes, size selection of a subset of the restriction fragments, PCR amplification and high-throughput sequencing of the size-selected fragments (Andrews et al., 2016). Typically, this technique has been used for rapid single nucleotide polymorphism (SNP) discovery and genotyping for genetic mapping of large populations in a variety of organisms (Baird et al., 2008; Baxter et al., 2011; Elshire et al., 2011; Pfender et al., 2011).

Since RAD-seq possesses several facilitative advantages, it has shown great promise and been broadly used to resolve phylogenetic relationships (Baird et al., 2008; Baxter et al., 2011; Henning et al., 2014; Pfender et al., 2011). First, RAD-seq can create a reduced representation of the genome, allowing detection of numerous informative SNPs. These markers provide unprecedented resolution of the framework phylogenies for several complex biological taxa, especially at intergeneric, interspecific and intraspecific taxonomic levels, such as Arundinarieae of Graminaeae (Wang et al., 2017), American oak clade (Hipp et al., 2014, 2018), Carex (Escudero et al., 2014; Massatti et al., 2016), *Pedicularis* section *Cyathophora* (Eaton and Ree, 2013), and *Primula tibetica* (Ren et al., 2017). Second, this approach serves as a powerful tool that has largely reduced the limitations of phylogenetic

### Table 1

| Demand for plant materials | Genome skimming | RNA-seq | Hyb-seq |
|---------------------------|-----------------|---------|---------|
| Fresh, silica-gel dried plant tissues and specimen | Fresh plant tissues | Fresh, silica-gel dried plant tissues and specimen |
| Low | High | Low |
| Yes | No | Yes |

| Material for sequencing Genome data obtained | Total genomic DNA | Complete plastid genome, nrDNA, partial mitochondrial genome, coding and non-coding genes |
|-----------------------------------------------|------------------|--------------------------------------------------------------------------------------------------|
| Total genomic DNA | cDNA | Randomly sequenced loci of vast majority of nuclear genome; coding genes |
| Restriction fragments | (SNP) mainly from nuclear genome; coding and non-coding genes |

| Targeted loci sequenced | Yes | No |
|-------------------------|-----|----|
| Identification of orthologs | Easy | Relatively easy |
| Missing data among species | No | Yes |

| Taxonomic levels for phylogenetic relationships | All levels from shallow to deep |
|-------------------------------------------------|------------------------------|
| Demands for plant materials Fresh, silica-gel dried plant tissues and specimen | Fresh plant tissues |
| Low | High |
| Yes | No |

| Applicable to specimen | Low |
|------------------------|-----|
| Material for sequencing Genome data obtained | Total genomic DNA |
| Nuclear gene conversion | cDNA |
| Nuclear genes were used to build a conformation of land plant has also been investigated through phylotranscriptomic (including 1KP data) analysis of 852 nuclear genes and 1,701,170 aligned sites (Wickett et al., 2014). However, transcriptomics requires living tissue for RNA extraction, and thus many existing tissue collections are unusable (Soltis et al., 2013). Therefore, a substantial amount of effort in future studies will be placed on resampling. In addition, RNA should be sampled from the same types of tissue and from individuals at the same life-history stage to obtain as many orthologous loci across samples as possible (Leemans and Leemans, 2013).

### 2.3. Phylogenomics using restriction site-associated DNA sequencing

Restriction site-associated DNA sequencing (RAD-Seq) and its related methods rely on the conservation of enzyme recognition sites (e.g., GBS: Genotyping-By-Sequencing; SLAF-seq: Specific-Locus Amplified Fragment Sequencing; SBG: Sequence-Based Genotyping). This technique can be used to survey hundreds or thousands of unlinked genetic markers adjacent to restriction sites from the nuclear genome (Baird et al., 2008; Elshire et al., 2011; Peterson et al., 2012). This approach mainly involves digestion of genomic DNA samples with restriction enzymes, size selection of a subset of the restriction fragments, PCR amplification and high-throughput sequencing of the size-selected fragments (Andrews et al., 2016). Typically, this technique has been used for rapid single nucleotide polymorphism (SNP) discovery and genotyping for genetic mapping of large populations in a variety of organisms (Baird et al., 2008; Baxter et al., 2011; Elshire et al., 2011; Pfender et al., 2011).

Since RAD-seq possesses several facilitative advantages, it has shown great promise and been broadly used to resolve phylogenetic relationships (Baird et al., 2008; Baxter et al., 2011; Henning et al., 2014; Pfender et al., 2011). First, RAD-seq can create a reduced representation of the genome, allowing detection of numerous informative SNPs. These markers provide unprecedented resolution of the framework phylogenies for several complex biological taxa, especially at intergeneric, interspecific and intraspecific taxonomic levels, such as Arundinarieae of Graminaeae (Wang et al., 2017), American oak clade (Hipp et al., 2014, 2018), Carex (Escudero et al., 2014; Massatti et al., 2016), *Pedicularis* section *Cyathophora* (Eaton and Ree, 2013), and *Primula tibetica* (Ren et al., 2017). Second, this approach serves as a powerful tool that has largely reduced the limitations of phylogenetic
reconstruction for non-genome species (Cariou et al., 2013; Etter and Johnson, 2012). Third, this high-throughput approach is relatively cost- and time-effective compared to other approaches such as whole genome sequencing (WGS) (Yang et al., 2016).

Although RAD-seq provides a potential approach to resolve shallow evolutionary timescales and phylogenetic relationships, its utility in resolving deep phylogenetic relationships is limited (Jones and Good, 2016). RAD-seq is mostly used for phylogenetic inference below the genus level (e.g., Eaton and Ree, 2013; Eaton et al., 2017; Emerson et al., 2010). However, RAD-seq data seem to be unreliable beyond moderate phylogenetic inference because restriction cut sites are usually not conserved across distinct taxa, leading to high levels of missing data and incorrect topologies (Leach et al., 2015; Rubin et al., 2012; Wagner et al., 2013). In addition, RAD-seq methods introduce several unique potential sources of error and bias in RAD-seq methods when doing phylogenetic studies. First, selecting a fragment size range by manual excision is subject to some human error, potentially leading to fewer orthologous fragments across individuals (McCormack et al., 2013b). Second, because the RAD tags are short (<300bp) and not targeted (Andrews et al., 2016), coverage can be difficult to estimate and the identification of orthologous fragments may vary among different assembly methods (Wang et al., 2017). Third, RAD-seq data notably suffer from a potentially large amount of missing data, which has triggered a serious debate on the effect of such missing data in reconstructing phylogenetic relationships (Eaton et al., 2017; Huang and Knowles, 2014; Leaché et al., 2015; Wagner et al., 2013). Another potential drawback is that most methods based on restriction digests are geared toward SNP generation, which may not be ideal for phylogenetic construction (Jones and Good, 2016).

2.4. Phylogenomics using targeted-capture

Targeted capture (targeted enrichment, hybridization enrichment, Hyb-Seq: combination of target enrichment and genome skimming approaches) is a technique that uses a hybridization reaction involving custom-designed short DNA or RNA probes in solution or on an array to capture thousands of target loci with sequences similar to the set of probes from fragmented genomic DNA libraries (Nicholls et al., 2015; Senapathy et al., 2010). Target capture, which allows simultaneous capture of low-copy nuclear genes and high-copy genomic targets (Weitemier et al., 2014), alleviates the limitations associated with DNA-seq and provides a prospective approach for plant phylogenetics. This high-throughput approach is more cost-effective at obtaining large data sets of orthologous loci across many individuals compared to WGS and multiplex PCR (Olson, 2007). A major challenge, however, is identifying the genomic sequences to be used for capture probes designed in non-reference species where a priori knowledge of target sequences is required (Elshire et al., 2011; Nicholls et al., 2015). This hurdle, nevertheless, can be resolved by utilizing de novo genomes for a large number of species that have already sequenced and the recent rapid accumulation of various kinds of genomic data (such as transcriptomes, genome skimming and RAD-seq). Such approaches have been successfully applied to resolve phylogenetic relationships of plants as well as other organisms (Brandley et al., 2015; Eytan et al., 2015; Fragoso-Martínez et al., 2017; Weitemier et al., 2014).

Targeted capture is appealing because it can theoretically be applied to any species with a de novo sequence assembly, (e.g., de novo whole genome sequencing, de novo RNA-seq transcriptomes or expressed sequence tag (EST) data), which have been or could be generated in the future (Bi et al., 2012, 2013). The flexibility of targeted capture for phylogenomic studies offers a tremendous advantage over other methods (Jones and Good, 2016). Customized capture designs from a set of probes with NGS can target 1) hundreds or thousands of orthologous loci (Faircloth et al., 2012; Mandel et al., 2014; Valderrama et al., 2018); 2) slowly or quickly evolving loci, including coding genes and non-coding regions (Lemmon and Lemmon, 2013; McCormack et al., 2013a); and 3) nuclear or organelle loci (Hedtke et al., 2013; Ilves and Lopez-Fernandez, 2014). Furthermore, targeted capture can work on a range of genomic DNA quality from fresh, silica-gel dried materials, even herbarium specimens of plants (Hart et al., 2016), whereas RNA-seq transcriptomes require fresh materials and RAD-seq requires high quality DNA isolated from fresh and silica-gel dried leaves of plants.

Selecting loci with appropriate evolutionary rates is very important when resolving a given relationship (Philippe et al., 2011). For example, deep phylogenetic nodes are resolved by using slowly evolving loci that retain signals of orthology across distant taxa (McCormack et al., 2012; Schott et al., 2017). High-throughput targeted capture of slowly evolving ultraconserved element (UCE) markers has been used to resolve deep nodes in the phylogenies of vertebrates (Crawford et al., 2015; Faircloth et al., 2012) as well as complex phylogenetic histories in flowering plants (Folk et al., 2017; Mandel et al., 2014). Capture of UCEs is appropriate for resolving relationships at all phylogenetic scales. Hutchinson et al. (2016) showed that a given set of loci can be identified from highly divergent reference genomes and can then be widely used for diverse taxa without continually redesigning custom probes (Jones and Good, 2016). In contrast, protein-coding sequences may be more suitable for plant phylogenetic reconstruction at moderate-to-deep evolutionary scales; these sequences are less conserved than UCEs, but more conserved than non-coding sequences (Mandel et al., 2014; Nicholls et al., 2015; Valderrama et al., 2018). To maximize phylogenetically informative sites at shallow evolutionary timescales (the inter-generic or intra-generic tips of the phylogeny), non-coding regions should be combined with coding regions, especially for recently radiated taxa. Phylogenetic studies based on target enrichment benefit from using a large number of putatively independent nuclear loci and their combination with plastid and mitochondrial genomes (Schmickl et al., 2016).

Targeted capture has emerged as a powerful approach in the genomic era to address plant phylogenetic questions and will be widely used for phylogenomics studies at various evolutionary scales. Hyb-Seq, a modified technique that combines target enrichment and genome skimming (Weitemier et al., 2014), has the advantages of both target enrichment and genome skimming, and will exhibit great promise for plant phylogenomics in the next few years.

3. Prospects in plant phylogenomics

Nuclear genomic data have been increasingly explored for phylogenomics. For example, nuclear repeat regions were used in Solanum section Lycopersicon to test the usefulness and power of phylogenomic analyses at inter- and intraspecific levels (Doddsworth et al., 2016). This approach provides additional evidence that might complement the results from organellar and the nuclear ribosomal cistron obtained from genome skimming. Currently, obtaining high-quality, low-cost genome sequences for a taxon of interest is routine. Third generation sequencing — real-time, single-molecule sequencing (PacBio and nanopore sequencing) — will be able to generate reads over 10 kb (or even 100 kb) (Bayley, 2015; Deamer et al., 2016; Eid et al., 2009; Shendure et al., 2017). Genome sequencing will benefit from the innovation of these sequencing technologies. Whole genome sequences have been used to infer the evolutionary relationships between closely related species of a flycatcher species complex
(Ficedula, Muscicapidae) and the early branches in the tree of life of modern birds and mammals (Jarvis et al., 2014; Nater et al., 2015; Sims et al., 2009). Resequencing, or mapping sequence reads to a reference genome to identify genetic variants, is less time-consuming than genome assembly (Shendure et al., 2017). It has been used to infer the geographic origin and migration history of the brown rat (Rattus norvegicus) (Zeng et al., 2017a), the genetic diversity of European ash trees (Fraxinus excelsior) (Sollars et al., 2016), and the history of apple domestication along the Silk Road (Duan et al., 2017). These achievements represent a prospective picture for resolving intractable plant relationships using whole genome and genome resequencing. Notably, the China National GeneBank and BGI-Shenzhen will lead the 10KP (10,000 Plants) Genome Sequencing Project, which will sequence and characterize representative genomes from every major clade of embryophyte, green algae, and protist (excluding fungi) within the next five years (Cheng et al., 2018). Once the project is finished, it will provide valuable genome resources for addressing numerous fundamental questions in evolutionary and comparative genomics and will greatly benefit our understanding of plant evolution and diversity. With the improvement and innovation of statistical and computational abilities for genome-sequence-based phylogenetics (Mirarab et al., 2014; Stamatakis et al., 2012), phylogenomic embraces a new era of discovery based on large amounts of genomic data and powerful analytical methods. The merger of phylogenomics with other biological disciplines, such as biogeography and ecology, will greatly advance our understanding of the origins and evolution of earth’s biodiversity.

Acknowledgements

We are grateful to Profs De-Zhu Li, Jun-Bo Yang and Hong-Tao Li for their discussion of the early version of the manuscript. We also thank Prof. Kevin S. Burgess from Columbus State University for improving the English of the manuscript. This study was supported by the Large-scale Scientific Facilities of the Chinese Academy of Sciences (Grant No: 2017-LSFGBOWS-01), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31000000), and the Program of Science and Technology Talents Training of Yunnan Province (2017HA014).

References

Andrews, K.R., Good, J.M., Miller, M.B., et al., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17, 81. Baird, N.A., Etter, P.D., Atwood, T.S., et al., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3, e3376. Baker, F.T., Leti, D., Yu, J., et al., 2015. Herbarium genomes: plastome sequence assembly from a range of herbarium specimens using an iterative Organellar Genome Assembly pipeline. Biol. J. Linn. Soc. Barrett, C.F., Davis, J.L., Leebens-Mack, J., et al., 2013. Plastid genomes and deep relationships among the cunemillid monocot angiosperms. Cladistics 29, 65–87. Barrett, C.F., Baker, W.J., Corner, J.R., et al., 2016. Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other cunemillid monocots. New Phytol. 209, 855–870. Baxter, S.W., Davy, J.W., Johnston, J.S., et al., 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. PLoS One 6, e19379. Bayley, H., 2015. Nanopore sequencing: from imagination to reality. Clin. Chem. 61, 25–31. Bellot, S., Renner, S.S., 2015. The plastomes of two species of the endoparasite genus Pilosiphon (Apodiformes: Apodidae) each retain just five or six possibly functional genes. Genome Biol. Evol. 8, 189–201. Bi, K., Vanderpool, D., Singhal, S., et al., 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genom. 13. Bi, K., Linderoth, T., Vanderpool, D., et al., 2013. Unlocking the vault: next-generation market place for biodiversity. Mol. Ecol. 22, E019–E032.
Valderrama, E., Richardson, J.E., Kidner, C.A., et al., 2018. Transcriptome mining for phylogenetic markers in a recently radiated genus of tropical plants (*Renealmia* L.f., Zingiberaceae). Mol. Phylogenet. Evol. 119, 13–24.

Vargas, O.M., Ortiz, E.M., Simpson, B.B., 2017. Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: asterreae: *Diplostephium*). New Phytol. 214, 1736–1750.

Wagner, C.E., Keller, I., Wittwer, S., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.

Wang, X., Ye, X., Zhao, L., et al., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (*Poaceae: bambusoideae*). Sci. Rep. 7, 11546.

Wagner, C.E., Keller, I., Wittwer, S., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.

Wagner, C.E., Keller, I., Wittwer, S., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.

Wang, X., Ye, X., Zhao, L., et al., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (*Poaceae: bambusoideae*). Sci. Rep. 7, 11546.

Wagner, C.E., Keller, I., Wittwer, S., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.

Wang, X., Ye, X., Zhao, L., et al., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (*Poaceae: bambusoideae*). Sci. Rep. 7, 11546.

Wagner, C.E., Keller, I., Wittwer, S., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.

Wang, X., Ye, X., Zhao, L., et al., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (*Poaceae: bambusoideae*). Sci. Rep. 7, 11546.

Wagner, C.E., Keller, I., Wittwer, S., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.

Wang, X., Ye, X., Zhao, L., et al., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (*Poaceae: bambusoideae*). Sci. Rep. 7, 11546.

Wagner, C.E., Keller, I., Wittwer, S., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.

Wang, X., Ye, X., Zhao, L., et al., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (*Poaceae: bambusoideae*). Sci. Rep. 7, 11546.

Wagner, C.E., Keller, I., Wittwer, S., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.

Wang, X., Ye, X., Zhao, L., et al., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (*Poaceae: bambusoideae*). Sci. Rep. 7, 11546.

Wagner, C.E., Keller, I., Wittwer, S., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.

Wang, X., Ye, X., Zhao, L., et al., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (*Poaceae: bambusoideae*). Sci. Rep. 7, 11546.

Wagner, C.E., Keller, I., Wittwer, S., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.

Wang, X., Ye, X., Zhao, L., et al., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (*Poaceae: bambusoideae*). Sci. Rep. 7, 11546.

Wagner, C.E., Keller, I., Wittwer, S., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.

Wang, X., Ye, X., Zhao, L., et al., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (*Poaceae: bambusoideae*). Sci. Rep. 7, 11546.

Wagner, C.E., Keller, I., Wittwer, S., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.

Wang, X., Ye, X., Zhao, L., et al., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (*Poaceae: bambusoideae*). Sci. Rep. 7, 11546.

Wagner, C.E., Keller, I., Wittwer, S., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.

Wang, X., Ye, X., Zhao, L., et al., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (*Poaceae: bambusoideae*). Sci. Rep. 7, 11546.