Molecular Phylogenetics and Evolution 128 (2018) 69–87

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe Cardueae (Compositae)

Sonia Herrando-Moraira[a,*], The Cardueae Radiations Group (Juan Antonio Calleja[b], Pau Carnicero[b], Kazumi Fujikawa[c], Mercè Galbany-Casals[b], Núria Garcia-Jacas[a], Hyoung-Tak Im[d], Seung-Chul Kim[e], Jian-Quan Liu[f], Javier López-Alvarado[b], Jordi López-Pujol[a], Jennifer R. Mandel[g], Sergi Massó[a], Iraj Mehregan[h], Noemí Montes-Moreno[a], Elizaveta Pyak[i], Cristina Roquet[j], Llorenç Sáez[b], Alexander Sennikov[k,l], Alfonso Susanna[a], Roser Vilatersana[a])

[a] Botanic Institute of Barcelona (IBB, CSIC-ICUB), Pg. del Migdia, s.n., 08038 Barcelona, Spain
[b] Systematics and Evolution of Vascular Plants (UAB) – Associated Unit to CSIC, Departament de Biologia Animal, Biologia Vegetal i Ecologia, Facultat de Biociències, Universitat Autònoma de Barcelona, ES-08193 Bellaterra, Spain
[c] Kochi Prefectural Makino Botanical Garden, 4200-6, Godaisan, Kochi 781-8125, Japan
[a] Botanic Institute of Barcelona (IBB, CSIC-ICUB), Pg. del Migdia, s.n., 08038 Barcelona, Spain
[d] Department of Biology, Chonnam National University, Gwangju 500-757, Republic of Korea
[e] Department of Biological Sciences, Sungkyunkwan University, Gyeonggi-do 440-746, Republic of Korea
[f] Key Laboratory for Bio-Resources and Eco-Environment, College of Life Sciences, Sichuan University, Chengdu, China
[g] Department of Biological Sciences, University of Memphis, Memphis, TN 38152, USA
[h] Department of Biology, Science and Research Branch, Islamic Azad University, Tehran, Iran
[i] Department of Botany, Inst. of Biology, Tomsk State University, RU-634050 Tomsk, Russia
[j] Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA (Laboratoire d'Ecologie Alpine), FR-38000 Grenoble, France
[k] Botanical Museum, Finnish Museum of Natural History, PO Box 7, FI-00014 University of Helsinki, Finland
[l] Herbarium, Komarov Botanical Institute of Russian Academy of Sciences, Prof. Popov str. 2, 197376 St. Petersburg, Russia

ARTICLE INFO

Keywords:
Asteraceae
COS targets
HybPiper
NGS filtering strategies
Phylogenetic noise
PHYLUCE

ABSTRACT

Target enrichment is a cost-effective sequencing technique that holds promise for elucidating evolutionary relationships in fast-evolving lineages. However, potential biases and impact of bioinformatic sequence treatments in phylogenetic inference have not been thoroughly explored yet. Here, we investigate this issue with an ultimate goal to shed light into a highly diversified group of Compositae (Asteraceae) constituted by four main genera: *Arctium*, *Cousinia*, *Saussurea*, and *Jurinea*. Specifically, we compared sequence data extraction methods implemented in two easy-to-use workflows, PHYLUCE and HybPiper, and assessed the impact of two filtering practices intended to reduce phylogenetic noise. In addition, we compared two phylogenetic inference methods: (1) the concatenation approach, in which all loci were concatenated in a supermatrix; and (2) the coalescence approach, in which gene trees were produced independently and then used to construct a species tree under coalescence assumptions. Here we confirm the usefulness of the set of 1061 COS targets (a nuclear conserved orthology loci set developed for the Compositae) across a variety of taxonomic levels. Intergeneric relationships were completely resolved: there are two sister groups, *Arctium-Cousinia* and *Saussurea-Jurinea*, which are in agreement with a morphological hypothesis. Intrageneric relationships among species of *Arctium*, *Cousinia*, and *Saussurea* are also well defined. Conversely, conflicting species relationships remain for *Jurinea*. Methodological choices significantly affected phylogenies in terms of topology, branch length, and support. Across all analyses, the phylogeny obtained using HybPiper and the strictest scheme of removing fast-evolving sites was estimated as the optimal. Regarding methodological choices, we conclude that: (1) trees obtained under the coalescence approach are topologically more congruent between them than those inferred using the concatenation approach; (2) refining treatments only improved support values under the concatenation approach; and (3) branch support values are maximized when fast-evolving sites are removed in the concatenation approach, and when a higher number of loci is analyzed in the coalescence approach.

* Corresponding author at: Botanic Institute of Barcelona (IBB, CSIC-ICUB), Pg. del Migdia, s.n., ES-08038 Barcelona, Spain.
E-mail address: sonia.herrando@gmail.com (S. Herrando-Moraira).

https://doi.org/10.1016/j.ympev.2018.07.012
Received 14 April 2018; Received in revised form 13 July 2018; Accepted 14 July 2018
Available online 21 July 2018

1. Introduction

1.1. Target enrichment strategies

The advent of the "target/hybrid enrichment" or "sequence capture" method has emerged in the last years as one of the most useful techniques in the field of phylogenomics and evolutionary studies (Cronn et al., 2012; Grover et al., 2012; Mamanova et al., 2009). This approach has provided significant advances, shedding light on previously unresolved evolutionary lineages analyzed using Sanger sequencing (Nicholls et al., 2015). This next generation sequencing (NGS) tool allows the recovery of hundreds to thousands of genetic markers from specific regions of the genome, even from degraded and ancient samples (Cronn et al., 2012). Remarkable advantages of this technique are: its reasonable sequencing cost, its power to resolve relationships at different taxonomic levels, and its reduced bioinformatic complexity compared to whole genome sequencing (Lemmon and Lemmon, 2013). The target DNA regions are enriched using probes or "baits". These can be specifically designed for the group of study via a known genome or transcriptome of a closely related species (e.g. Folk et al., 2015; García et al., 2017; Schmickl et al., 2016; Syring et al., 2016), or universally conserved loci (e.g., anchored hybrid enrichment, AHE) as for vertebrates (Lemmon et al., 2012) or angiosperms (Buddenhagen et al., 2016).

Concerning the Compositae or Asteraceae (both terms used to refer to the sunflower family; hereafter Compositae), Mandel et al. (2014) recently developed a target enrichment method, which uses the Hyb-Seq (sequence capture) approach (Weitemier et al., 2014), comprising a probe set of 9678 baits targeting a total of 1061 conserved orthology loci (hereafter COS) in this family. These COS loci were identified from thousands of expressed sequence tags (EST) across three available genomes of the family (see Mandel et al., 2014). This method has already proven useful at varied taxonomical scales, from deep Compositae nodes to shallower ones (Mandel et al., 2014, 2015, 2017; Siniscalchi et al., in preparation). In addition, the method allows the recovery of plastome data captured from off-target sequenced reads (Mandel et al., 2015). Nevertheless, the analytical power of this approach to resolve species relationships of recently and rapidly radiated genera in the family remains untested. In addition, the above cited previous works using the Compositae COS targets (Mandel et al., 2014, 2015, 2017) were performed following only one bioinformatics workflow for target sequences extraction, called PHYLUCE (Faircloth, 2015).

The last point seems crucial, since it has not been thoroughly investigated yet whether different bioinformatics extraction approaches yield congruent phylogenetic results, and whether these methodological choices could lead to bias in phylogenetic reconstruction. In recent years, a great number of easy-to-use workflows and automated pipelines are emerging to be used as target extraction procedures. The pipeline PHYLUCE (Faircloth, 2015) was initially designed for ultra-conserved elements (UCEs, Faircloth et al., 2012) and applied to a wide range of animal groups: birds (Hosner et al., 2015; Moyle et al., 2016), skinks (Bryson et al., 2017), ants (Ješovnik et al., 2017), and fishes (Burress et al., 2017; Longo et al., 2017). A bioinformatic approach for AHE was proposed in Prum et al. (2015) and used in several plant studies (Buddenhagen et al., 2016; Fragoso-Martínez et al., 2017; Mitchell et al., 2017; Wanke et al., 2017). Another method, HybPiper (Johnson et al., 2016) was designed specifically for Hyb-Seq data, implementing the ability to target exons and introns separately. The HybPiper workflow also offers the option to identify and separate paralogous copies. HybPiper has already been successfully applied to analyse data from captured target loci in plants (e.g. Crowl et al., 2017; Landis et al., 2017; Chau et al., 2018; Gernandt et al., 2018; Kates et al., 2018; Medina et al., 2018; Stubbs et al., 2018; Vatanparast et al., 2018). Other new and promising tools are aTRAM (Allen et al., 2015, 2017), HybPhyloMarker (Fér and Schmickl, 2018), and SECAPR (Andermann et al., 2018). From these published pipelines, we selected for this study two of the most commonly used approaches, PHYLUCE and HybPiper, to explore the technical differences between them and asses the consequences in inferred phylogenies of choosing one or another.

1.2. Parsing phylogenetic signal from noise in NGS studies

Despite the large amount of DNA sequence characters generated with NGS, the true gene genealogy can be obscured by various kinds of "phylogenetic noise" (Straub et al., 2014; Townsend et al., 2012). Potential sources of noise in nucleotide sequences include unusually fast-evolving sites, rich-indel regions, and ambiguous sequence calls, which may lead to substitution saturation, i.e. convergence in nucleotide states (homoplasy) that contradicts the real phylogenetic signal, and bias the ancestry character-state reconstructions (Rokas and Carroll, 2006). Additional noise may accumulate in all study phases due to sequencing errors, inaccurate assembly, or incorrect orthology assignment. Another possible source of error that should be taken into account with NGS data is the incorrect allele phasing in polyploid systems (Eriksson et al., 2018), in which phylogenetic trees can be often reconstructed from consensus sequences or chimeric consensus sequences rather than real allele sequences (Kates et al., 2018). Consequences of ignoring possible phylogenetic noise are well documented (Kostka et al., 2008; Straub et al., 2014; Townsend et al., 2012), and may lead to long-branch attraction artifacts, topological differences among alternative reconstructions, or high support values for erroneous relationships (Dornburg et al., 2014; Jeffroy et al., 2006; Salichos and Rokas, 2013).

Part of this phylogenetic noise can be reduced with standard practices such as cleaning raw reads by filtering based on quality scores and alignment trimming (i.e. removal of ambiguously aligned and indel-rich positions). However, final trimmed alignments commonly used to perform phylogenetic inferences may still contain considerable levels of noise. Currently, standard procedures to deal with this issue are not well established, and we still lack a widely applicable refining metric to minimize the negative effects of phylogenetic noise and maximize the likelihood of an accurate phylogenetic reconstruction. Many recent studies based on target enrichment incorporate diverse filtering strategies at different components of data matrices, such as species, positions, or even entire sets of loci (see Table 1). Among all these practices, the most commonly used is the exclusion of loci recovered for a low number of species, which aims to reduce the effects of missing data and systematic bias on tree inference (see Hosner et al., 2015 for further details on potential impacts of missing data).

1.3. Resolving radiations and the case of the groups Arctium-Cousinia and Saussurea-Jurinea (tribe Cardueae)

Explosive diversification events (referred here as radiations) represent events in which many species or lineages evolved from a common ancestor in a short time period (Wen et al., 2013, 2014), caused by geographic isolation, dispersal barriers, sexual selection, or in some cases by ecological divergence or acquisition of novel key traits (Givnish, 2015). These events may leave few genomic traces, yielding few nucleotide differences among species derived from a common radiation, and thus hindering the reconstruction of phylogenetic relationships among them. As a consequence, unresolved phylogenies with short internal branches or large polytomies have been often recovered with traditional Sanger sequenced markers in recently diverged genera, hampering the in-depth study of radiations. With the emergence of NGS techniques, research focused on plant radiations are significantly increasing (*Heuchera* L., Folk et al., 2015; *Inga* Mill., Nicholls et al., 2015; Cariceae-Dulichieae-Scirpeae clade in Cyperaceae Juss., Léveillé-Bourret et al., 2016; order Zingiberales Griseb., Sass et al., 2016; *Salvia* L. subgenus *Calosphace* (Benth.) Epling, Fragoso-Martínez et al., 2017; *Protea* L., Mitchell et al., 2017; *Aristolochia* L., Wanke et al., 2017; "*Adenocalymma-Neojobertia*" clade from