Molecular Phylogenetics and Evolution 119 (2018) 13–24

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Transcriptome mining for phylogenetic markers in a recently radiated genus of tropical plants (*Renealmia* L.f., Zingiberaceae)

Eugenio Valderrama[a,b,c,*], James E. Richardson[a,d], Catherine A. Kidner[a,e], Santiago Madriñán[f], Graham N. Stone[a,b]

[a] Royal Botanic Garden Edinburgh, Edinburgh, Scotland, United Kingdom
[b] Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, Scotland, United Kingdom
[c] Laboratorio de Análisis del Comportamiento Humano, Facultad de Psicología y Grupo de Investigación de Biología (GRIB), Universidad El Bosque, Bogotá, Colombia
[d] Programa de Biología, Universidad del Rosario, Bogotá, Colombia
[e] Institute of Molecular Plant Sciences, University of Edinburgh, Edinburgh, Scotland, United Kingdom
[f] Laboratorio de Botánica y Sistemática, Universidad de Los Andes, Bogotá, Colombia

ARTICLE INFO

Keywords:
Phylogenetics
Transcriptomes
Markers
Diversification
Zingiberaceae
Neotropical

ABSTRACT

The reconstruction of relationships within species-rich groups that have recently evolved in biodiversity hotspots is hampered by a lack of phylogenetically informative markers. It is also made difficult by the lack of sampling necessary to reconstruct a species-level phylogeny. We use transcriptome mining to search for markers to reconstruct a phylogeny of the amphi-Atlantic genus *Renealmia* L. f. (Zingiberaceae). We recover seven introns from single copy genes and use them to reconstruct the phylogeny of the genus together with a commonly used phylogenetic marker, internal transcribed spacers of ribosomal DNA (ITS) that has previously been used to reconstruct the phylogeny of the genus. We targeted genes with low numbers of base pairs that improves sequencing success using highly degraded DNA from herbarium specimens. The use of herbarium specimens greatly increased the number of species in the study as these were readily available in historical collections. Data were obtained for 14 of the 17 African species and 54 of the 65 Neotropical species. The phylogeny was well-supported for a number of *Renealmia* subgroups although relationships among those clades remained poorly supported.

1. Introduction

Recent diversification of the species-rich Neotropical flora presents a challenge for studies attempting to elucidate evolutionary history at the species level. Widely used phylogenetic markers struggle to recover species level relationships within recent and species-rich lineages (Hughes and Eastwood, 2006; Nicholls et al., 2015; e.g. Richardson et al., 2001; Särkinen et al., 2007) a scenario that could be expected when different and non-mutually exclusive processes affect radiations. Recovering phylogenies when lineages are recent and so lack lineage specific mutations is challenging, also when influenced by incomplete lineage sorting because they are recent and/or comprise large population sizes and finally when introgression through hybridization has occurred within them (Maddison, 1997; Maddison and Knowles, 2006; Pamilo and Nei, 1988). Although this pattern points to the importance of these processes, it is necessary to improve our understanding of the evolutionary history of these lineages in order to elucidate the reasons for such diversifications and gain insights into the mechanisms that yielded the outstanding diversity of the Neotropics.

The traditional approach over the last two decades to obtain species-level phylogenies in plants has been to amplify a limited set of non-coding cpDNA loci (Shaw et al., 2007) and the nrDNA internal transcribed spacer (ITS) region (e.g. Armstrong et al., 2014; Álvarez and Wendel, 2003). These are expected to be variable at low taxonomic levels and are easy to sequence using universal PCR primers. This approach has shed light on many relationships among genera and even at the species level, but has also led to poorly resolved phylogenies that are not very informative for studies beyond classification purposes, even in large and morphologically diverse genera (Hughes et al., 2006). In recent radiations, incomplete lineage sorting and low differentiation among species are expected to cause incongruence between the phylogenetic signals of sampled loci, necessitating approaches that incorporate coalescent variation. During the last decade, several methods have been developed that accommodate incongruence between gene

* Corresponding author at: Facultad de Psicología y Programa de Biología, Universidad El Bosque, Bogotá, Colombia.
E-mail addresses: e.valderrama.e@gmail.com (E. Valderrama), jamese.richardson@urosario.edu.co (J.E. Richardson), c.kidner@rbge.ac.uk (C.A. Kidner), samadrin@uniandes.edu.co (S. Madriñán), graham.stone@ed.ac.uk (G.N. Stone).

http://dx.doi.org/10.1016/j.ympev.2017.10.001
Received 27 June 2016; Received in revised form 17 July 2017; Accepted 3 October 2017
Available online 04 October 2017
1055-7903/

trees and the underlying species tree (Heled and Drummond, 2010; Liu, 2008; Mirarab et al., 2014; Shaw et al., 2013). However, to estimate the species trees it is necessary to sample multiple low-copy number, phylogenetically-informative, orthologous and unlinked loci across the genome of the study lineage, a challenging task in some plant groups where traditional markers are not informative enough (Hughes et al., 2006).

Recent developments in DNA sequencing technology (Metzker, 2009) have facilitated the generation of genome-wide phylogenetic datasets, allowing resolution of the evolutionary history of challenging study systems (Brassac and Blattner, 2015; e.g. Nicholls et al., 2015; Yang et al., 2015). An alternative approach to genome-wide sampling of lineages with large genomes that avoids the costs of either generating high enough coverage to reliably sequence homologous regions in multiple accessions is to sequence transcribed sequences of the genome (the transcriptome) with RNA extracted from living specimens (Wang et al., 2009). Transcriptomes represent a substantial reduction of the genome that can be sequenced with a smaller budget. The development of specific software for assembling transcripts (Haas et al., 2013) and the growing number of lineages for which transcriptomes are becoming available in public repositories (Droc et al., 2013; Goodstein et al., 2012) make *de-novo* transcriptomes easier to assemble and compare. Transcriptomes can be used as a genome-wide sample to target useful markers to estimate phylogenies (Chamala et al., 2015; Rothfels et al., 2013) and could be used for selecting markers for targeted enrichment (e.g. Nicholls et al., 2015).

Besides using appropriate markers, a common problem in phylogenetic studies of widespread groups is that some species are only available for sampling as herbarium specimens. Sampling widespread species from throughout their distributions or local endemics in the field to acquire high quality tissue samples would require considerable research resources, intricate logistics and long-term studies to estimate species level phylogenies for species-rich groups. The issue is exacerbated in the Neotropics by the widespread habit of using alcohol to preserve specimens in difficult environments until drying facilities (*e.g.* the local herbarium) are reached by the botanists, reducing the already diminished quality of the DNA that can be extracted from herbarium specimens (Särkinen et al., 2012). However, the likelihood of successful PCR amplifications of DNA extracted from herbarium specimens can be increased by modifying the standard DNA extraction protocols and targeting short amplicons (Särkinen et al., 2012). Field collection of RNA samples for transcriptomic approaches is more challenging, requiring more complex preservatives (such as RNA-later) and cryostorage – and RNA preservation is usually very poor in herbarium material.

In this study we present an alternative method to generate species level phylogenies of recent radiations of plants with big genomes and for which several species are only available as herbarium specimens. We generated three transcriptomes and compared them to available public repositories to target short introns within potentially conserved, orthologous and low copy-number loci. Those loci were amplified with PCR's that can be used on low quality DNA extractions from herbarium specimens to obtain markers to cover as much species as possible of the genus *Renealmia* L.f. (Zingiberaceae) to estimate a species-level phylogeny. Finally, we use the resulting phylogenies to ask whether groups and subgroups defined by morphological characters (see Supplementary Notes 1) are supported by molecular data, and whether some widespread species are resolved as monophyletic clades.

2. Material and methods

2.1. de novo transcriptomes

2.1.1. Tissue

Three transcriptomes were generated from individuals held in the living collection of the Royal Botanic Garden Edinburgh under the accession numbers 20080229A, 19740104 and 19750180 corresponding to the Neotropical species *R. alpinia* (Rottb.) Maas, the African species *R. battenbergiana* H.A. Cummins ex Baker and a species of *Alpinia* Roxb., respectively (Supplementary Table 1).

2.1.2. RNA extraction

Young buds were sampled so that a variety of tissues including young leaves, rhizomes and meristems could be incorporated to harvest mRNAs for as many as possible of the expressed proteins in the vegetative parts. The buds were frozen immediately after collection in liquid nitrogen and kept frozen during the grinding process. PureLink Plant RNA Reagent (Life Technologies) was used following the manufacturer's protocol for small scale RNA isolation with two adjustments. 500 µL of acid phenol:chloroform (5:1, pH 4.7; SIGMA-ALDRICH) was added to the aqueous phase obtained after the protocol's first chloroform addition, then mixed and centrifuged at 12,000*g* for ten minutes at 4 °C. The aqueous phase to which isopropyl alcohol was added was left for at least an hour at −20 °C to improve precipitation. Integrity of the extracted RNA was assessed on 1% agarose gels stained with ethidium bromide. RNA concentrations were measured using a Qubit 2.0 Fluorometer (Life Technologies) after applying the RQ1 DNase (Promega) treatment; both procedures followed manufacturer's protocols. Pooling of multiple extractions (up to four) was necessary to increase the total quantity of RNA per species.

2.1.3. de novo transcriptomes

TruSeq RNA Libraries (Illumina) were prepared and sequenced in a MiSeq v2 (Illumina) lane (12×10^6 base pairs [bp] in 150 bp paired-end-reads) by Edinburgh Genomics at the University of Edinburgh. *Trimmomatic* 0.3 (Bolger et al., 2014) was used to remove adapters and bases with poor quality using a sliding window trimming approach with windows of 4 bp and a threshold value for the average Phred score of 15. Phred scores are logarithmically related to the base-calling error probability, a Phred score of 15 represents a probability of incorrect base call of 0.0316 (Ewing and Green, 1998) and represents a coarse filter to avoid losing too much information at this step in the assemblies. Bases with Phred scores below three at both ends of the reads and reads shorter than 36 bp were removed in *Trimmomatic*. *FastQC* (Andrews et al., 2015) was used to check quality before and after trimming. The remaining reads after trimming were assembled in *Trinity* with default settings (Haas et al., 2013).

2.1.4. Choosing markers for phylogenetic analyses

To choose appropriate genes in the transcripts for phylogenetic analysis it is necessary to distinguish which genes evolved from a common ancestral gene by speciation (orthologous genes) from those that are related by duplication within a genome (paralogous genes). To estimate which genes in the transcripts of the three species are potentially orthologous we generated a list of the Reciprocal Best Hits (RBH) using *BLAST* (Altschul et al., 1990). We assumed that the transcripts of orthologous genes present in two different genomes will identify each other as reciprocal best hits due to shared common ancestry (Tatusov et al., 1997; Ward and Moreno-Hagelsieb, 2014).

We checked if the potential orthologous genes found in the three species share the same best *BLAST* hit in the transcriptomes of *Arabidopsis thaliana* (TAIR10; Lamesch et al., 2012) and *Musa acuminata* (Droc et al., 2013). We used the unique locus identifier of *A. thaliana* genes that were best *BLAST* hits to identify the potential orthologs among the three transcriptomes included in the COSII list. The COSII lists 2869 genes selected within the coding sequences (CDS) of *A. thaliana* and representatives of the euasterids that, according to phylogenies and reciprocal *BLAST* methods, are potentially orthologs and single-copy genes (F. Wu et al., 2006). We also used the *A. thaliana* locus identifiers to visually inspect the gene model of each potential orthologous gene in all the available data in *Phytozome* 9.1 for model angiosperms (Goodstein et al., 2012). Loci that included introns of