Molecular Phylogenetics and Evolution 157 (2021) 107068

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

# Phylogenomics of Gesneriaceae using targeted capture of nuclear genes

Ezgi Ogutcen [a], Camille Christe [a], Kanae Nishii [b,c], Nicolas Salamin [d], Michael Möller [b], Mathieu Perret [a,*]

[a] Conservatoire et Jardin botaniques de la Ville de Genève and Department of Botany and Plant Biology, University of Geneva, 1292 Chambésy, Switzerland
[b] Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, Scotland, UK
[c] Kanagawa University, 2946, Tsuchiya, Hiratsuka-shi, Kanagawa 259-1293, Japan
[d] Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland

ARTICLE INFO

*Keywords:*
African violet
Phylogenetics
Probe design
Sequence capture
Systematics
Target enrichment

ABSTRACT

Gesneriaceae (ca. 3400 species) is a pantropical plant family with a wide range of growth form and floral morphology that are associated with repeated adaptations to different environments and pollinators. Although Gesneriaceae systematics has been largely improved by the use of Sanger sequencing data, our understanding of the evolutionary history of the group is still far from complete due to the limited number of informative characters provided by this type of data. To overcome this limitation, we developed here a Gesneriaceae-specific gene capture kit targeting 830 single-copy loci (776,754 bp in total), including 279 genes from the Universal Angiosperms-353 kit. With an average of 557,600 reads and 87.8% gene recovery, our target capture was successful across the family Gesneriaceae and also in other families of Lamiales. From our bait set, we selected the most informative 418 loci to resolve phylogenetic relationships across the entire Gesneriaceae family using maximum likelihood and coalescent-based methods. Upon testing the phylogenetic performance of our baits on 78 taxa representing 20 out of 24 subtribes within the family, we showed that our data provided high support for the phylogenetic relationships among the major lineages, and were able to provide high resolution within more recent radiations. Overall, the molecular resources we developed here open new perspectives for the study of Gesneriaceae phylogeny at different taxonomical levels and the identification of the factors underlying the diversification of this plant group.

## 1. Introduction

The Gesneriaceae is a pantropical plant family of perennial herbs, shrubs, or small trees that comprises around 150 genera and over 3400 species (Weber et al., 2013; Möller et al., 2016). The colonization of a wide range of habitats and the evolution of specialized plant–animal interactions to achieve pollination and seed dispersal has strongly influenced the diversification of this clade since its origin around 70 million years ago (Roalson and Roberts, 2016; Serrano-Serrano et al., 2017). The extensive diversity of Gesneriaceae in habit and floral morphology coupled with high levels of convergence in these traits caused considerable confusion in the early taxonomy of this family (Jong and Burtt, 1975; Clark et al., 2012). To date, phylogenetic inference in this plant group has mainly relied on plastid markers (e.g., *atp*B-*rbc*L, *psb*A-*trn*H, *trn*L-trnF, *ndh*F) and few multi-copy nuclear ribosomal regions such as ITS, and to a lower extent, low-copy nuclear genes such as *GLUTAMINE SYNTHETASE* (*ncpGS*) and *CYCLOEDIA* (*CYC*) (reviewed in Möller and Clark, 2013; Roalson and Roberts, 2016). Phylogenetic hypotheses derived from these genetic markers provided the framework to redefine the generic and tribal boundaries and to develop a new formal classification of the family (Zimmer et al., 2002; Perret et al., 2003; Roalson et al., 2005; Clark et al., 2006; Möller et al., 2009, 2011; Clark et al., 2012; Weber et al., 2013). The analyses of these sequence data using supermatrix approaches also provided large scale phylogenetic hypotheses for the entire family (768 species; Roalson and Roberts, 2016) and the Gesnerioideae subfamily (583 species; Serrano-Serrano et al., 2017). However, the limited number of informative sites provided by these DNA regions currently hinders our understanding of the phylogenetic relationships within the most diverse genera such as *Besleria* (Clark et al., 2006), *Columnea* (Schulte et al., 2014), *Cyrtandra* (Atkins et al., 2019), and *Streptocarpus* (Nishii et al., 2015). In addition, the few available nuclear sequences (e.g. ITS, *ncpGS*, *CYC*) are highly divergent across the subfamilies and at higher ranks, thus preventing the use of these loci to resolve deep relationships within Gesneriaceae and

* Corresponding author.
*E-mail address:* Mathieu.perret@ville-ge.ch (M. Perret).

https://doi.org/10.1016/j.ympev.2021.107068
Received 27 May 2020; Received in revised form 30 December 2020; Accepted 4 January 2021
Available online 7 January 2021

**Table 1**
Voucher information for the 78 samples used in this study.

| Species | Accession ID | Lab ID | Voucher specimen; provenance |
|---|---|---|---|
| *Agalmyla chalmersii* (F.Muell.) B.L.Burtt | SAMN17001995 | P691 | ex Lae 252 (E); cult., RBGE, 19,661,971 |
| *Anetanthus gracilis* Hiern | SAMN17002070 | P907 | Peixoto, M. & Chautems, A. 28 (G); wild, Brazil |
| *Anna submontana* Pellegr. | SAMN17002003 | P700 | Möller, M. & Qi, Q 01–85 (E, PE); wild, China 2001 |
| *Besleria labiosa* Hanst. | SAMN17002047 | P883 | Wiehler & Steyermark 72,453 (E); cult., RBGE, 19,822,666 |
| *Besleria solanoides* Kunth | SAMN17002063 | P900 | Perret, M. et al. 288 (G); wild, Ecuador 2019 |
| *Chrysothemis friedrichsthaliana* (Hanst.) H.E.Moore | SAMN17002060 | P819 | Perret, M. & Chautems, A. 37 (G); cult., CJBG, 20160825 J0 |
| *Cobananthus calochlamys* (Donn.Sm.) Wiehler | SAMN17002043 | P762 | Perret, M. 354 (G); cult., CJBG, 20171586C |
| *Codonanthopsis elegans* (Wiehler) Chautems & Mat.Perret | SAMN17002033 | P748 | Perret, M. 355 (G); cult., CJBG, 20111623JO |
| *Columnea ulei* Mansf. | SAMN17002024 | P747 | Perret, M. 356 (G); cult., CJBG, 20111639JO |
| *Corallodiscus lanuginosus* (Wall. ex DC.) B.L.Burtt | SAMN17002023 | P707 | Möller, M. & Zhou, P. 10-1681B (XTBG); wild, China 2010 |
| *Corytoplectus speciosus* (Poepp.) Wiehler | SAMN17002009 | P885 | Leiden, B.G. 68 (E); cult., RBGE, 19540131*A |
| *Cremosperma hirsutissimum* Benth. | SAMN17002049 | P765 | Perret, M. et al. 205 (G); wild, Colombia 2016 |
| *Cremosperma nobile* C.V.Morton | SAMN17002035 | P899 | Clavijo, L. 2260 (CUVC); wild, Colombia 2019 |
| *Cyrtandra crockerella* Hilliard | SAMN17002015 | P710 | Mendum, M. 43 (E); cult., RBGE, 20,001,505 |
| *Cyrtandra oblongifolia* (Blume) C.B.Clarke | SAMN17002012 | P695 | MAP 5 (E); cult., RBGE, 19,912,412 |
| *Didissandra frutescens* (Jack) C.B.Clarke | SAMN17001999 | P901 | Rafidah, A.R. FRI 64,355 (KEP); wild, Malaysia |
| *Didymocarpus antirrhinoides* A.Weber | SAMN17002064 | P697 | Jong, K. 9009 (E); cult., RBGE, 19,650,167 |
| *Dorcoceras hygrometricum* Bunge | SAMN17002001 | P709 | Möller, M. & Gao, L.M. 05–686 (E); wild, China 2005 |
| *Drymonia serrulata* (Jacq.) Mart | SAMN17002016 | P736 | Perret, M. 357 (G); cult., CJBG, 20036936NO |
| *Episcia cupreata* (Hook.) Hanst. | SAMN17002034 | P763 | Perret, M. 358 (G); cult., CJBG, 20150058IO |
| *Epithema carnosum* Benth. | SAMN17002065 | P902 | Möller, M. & Wei, Y.G. 06-864b (E); wild, China 2006 |
| *Gasteranthus pansamalanus* (Donn.Sm.) Wiehler | SAMN17002022 | P898 | Perret, M. 259 (G); wild, Ecuador 2019 |
| *Gesneria ventricosa* Sw. | SAMN17002061 | P746 | Perret, M. 359 (G); cult., CJBG, 20111187 J1 |
| *Gloxinella lindeniana* (Regel) Roalson & Boggan | SAMN17002032 | P760 | Perret, M. 360 (G); cult., CJBG, 20160159JO |
| *Gloxinia perennis* (L.) Fritsch | SAMN17002019 | P742 | Perret, M. 361 (G); cult., CJBG, 20150663I1 |
| *Gyrocheilos retrotrichum* W.T.Wang | SAMN17002000 | P696 | Möller, M. & Wei, Y.G. 07–1136 (E); wild, China 2007 |
| *Haberlea rhodopensis* Friv. | SAMN17002053 | P890 | Sophia Univ. B. G. 972 (E); cult., RBGE, 19,281,002 |
| *Kohleria hirsuta (Kunth) Regel* | SAMN17002051 | P887 | Mason, L.M. 476 (E); cult., RBGE, 19,551,049 |
| *Leptoboea multiflora* subsp. *grandifolia* B.L.Burtt | SAMN17002005 | P703 | Middleton, D. J. 5680 (E); cult., RBGE, 20121416*A, MMOG 280 |
| *Litostigma crystallinum* Y.M.Shui & W.H.Chen | SAMN17002013 | P712 | Shui, Y.M. 43,865 (KUN); wild, China |
| *Loxostigma griffithii* (Wight) C.B.Clarke | SAMN17002057 | P894 | Kew/Edinburgh Kanchenjunga exp. (1989) 940 (E); cult., RBGE, 19,892,473 |
| *Microchirita prostrata* J.M.Li & Z.Xia | SAMN17002056 | P893 | Möller, M. & Gao, L.M. 06–945 (E); wild, China 2006 |
| *Napeanthus primulifolius* (Raddi) Sandwith | SAMN17002071 | P908 | Araujo, A.O. 470 (ESA); wild, Brazil |
| *Nautilocalyx bicolor* (Hook.) Wiehler | SAMN17002026 | P751 | Perret, M. 362 (G); cult., CJBG, 20110030JO |
| *Nematanthus monanthos* (Vell.) Chautems | SAMN17002044 | P840 | Chautems, A. & Perret, M. 08–601 (G); cult., CJBG, 20111667 J0 |
| *Oreocharis farreri* (W.G.Craib) M.Möller & A.Weber | SAMN17001996 | P692 | Zhou, P. 2010–020 (XTBG); wild, China 2010 |
| *Oreocharis magnidens* Chun ex K.Y.Pan | SAMN17001997 | P693 | Möller, M. & Wei, Y.G. 06–896 (E); wild, China 2006 |
| *Oreocharis sinensis* (Oliv.) M.Möller & A.Weber | SAMN17001998 | P694 | Möller, M. & Wei, Y.G. 09–1329 (E); wild, China 2009 |
| *Paliavana prasinata* (Ker Gawl.) Benth. | SAMN17001994 | P679 | Chautems, A. & Perret, M. 00–013 (G); cult., CJBG, 19,662,361 |
| *Paliavana tenuiflora* Mansf. | SAMN17002027 | P752 | Perret, M. 363 (G); cult., CJBG, 20036853 N1 |
| *Paraboea rufescens* (Franch.) B.L.Burtt | SAMN17002010 | P708 | Zhou, P. 2010-074A (XTBG); wild, China 2010 |
| *Primulina lutea* (Yan Liu & Y.G.Wei) Mich.Möller & A.Weber | SAMN17002002 | P698 | Möller, M. & Wei, Y.G. 06–909 (E); wild, China 2006 |
| *Ramonda myconi* (L.) Rchb. | SAMN17002058 | P895 | Perret, M. 364 (G); cult., CJB, s.n. |
| *Raphiocarpus petelotii* (Pellegr.) B.L. Burtt | SAMN17002004 | P701 | Goodwin, S. & Cherry, R. 92/208 (E) ; cult., RBGE, 19,982,405 |
| *Reldia minutiflora* (L.E. Skog) L.P. Kvist & L.E. Skog | SAMN17002037 | P769 | Perret, M. et al. 179 (G); wild, Colombia 2016 |
| *Rhabdothamnus solandri* A.Cunn. | SAMN17002048 | P884 | Kealy, J. s.n. (E); cult., RBGE, 19660192*A |
| *Rhynchoglossum gardneri* W.L.Theob. & Grupe | SAMN17002008 | P706 | Theobald, W. & Grupe 2309 (E); cult., RBGE, 19,682,727 |
| *Rhytidophyllum tomentosum* (L.) Mart. | SAMN17002050 | P886 | ex Hamburg B.G. 138; cult., RBGE, 19591512*A |
| *Sanango racemosum* (Ruiz & Pav.) Barringer | SAMN17002059 | P896 | Neill, D.A. 9458 (US); herbarium G |
| *Sarmienta repens* Ruiz & Pav. | SAMN17002014 | P713 | Gardner, M. & Knees, S. 4033 (E); cult., RBGE, 19,882,757 |
| *Seemannia sylvatica* (Kunth) Hanst. | SAMN17002025 | P750 | Araujo, A.O. et al. 603 (G); cult.,CJBG, 20151341JO |
| *Sinningia araneosa* Chautems | SAMN17002042 | P798 | Chautems, A. & Perret, M. 00–016 (G); cult., CJBG, 20131657I0 |
| *Sinningia bullata* Chautems & M.Peixoto | SAMN17002039 | P793 | Reis, A. et al. 5040 (G); cult., CJBG, 20131729I0 |
| *Sinningia cardinalis* (Lehm.) H.E.Moore | SAMN17002018 | P740 | Perret, M. 365 (G); cult., CJBG, 20150242IO |
| *Sinningia conspicua* (Seem.) G.Nicholson | SAMN17002041 | P797 | Chautems, A & Perret, M. 00–008 (G); cult., CJBG, 20111013 J0 |
| *Sinningia harleyi* Chautems | SAMN17002017 | P739 | Perret, M. 366 (G); cult., CJBG, 20161282NO |
| *Sinningia helioana* Chautems & Rossini | SAMN17002040 | P795 | Salviani & Peixoto (MBML); cult., CJBG, 20151354 J0 |
| *Sinningia schiffneri* Fritsch | SAMN17002007 | P705 | Chautems, A. & Perret, M. 97–010 (G); cult., CJBG, 19781514*A |
| *Sinningia* sp. *nov. 1* | SAMN17002021 | P744 | Perret, M. 367 (G); cult., CJBG, 201316641IO |
| *Sphaerorrhiza sarmentiana* (Gardner ex Hook.) Roalson & Boggan | SAMN17002036 | P767 | Araujo, A.O. et al. 539 (ESA); cult., CJBG, AC-3807 |
| *Streptocarpus cyaneus* S.Moore indiv.1 | SAMN17002066 | P903 | Hughes, M. et al. 1377 [MMOG 475]; cult., RBGE, 20,060,901 |
| *Streptocarpus cyaneus* S.Moore indiv.2 | SAMN17002067 | P904 | Scott, D. s.n. (E) [MMOG 476]; cult., RBGE, 19,911,951 |
| *Streptocarpus formosus* (Hilliard & B.L.Burtt) T.J.Edwards | SAMN17002068 | P905 | Burtt, B.L. 6063 (E) [MMOG 478]; cult., RBGE, 20,141,208 |
| *Streptocarpus glandulosissimus* Engl. | SAMN17002055 | P892 | Hilliard, O.M. 348 (E); cult., RBGE, 19652118*B |
| *Streptocarpus modestus* Britten | SAMN17002069 | P906 | Hughes, M. et al. MH1127 (E) [MMOG-477]; cult., RBGE, 20,120,811 |
| *Streptocarpus rexii* (Bowie ex Hook.) Lindl. | SAMN17002054 | P891 | Jong, K. 1226 (E); cult., RBGE, 20110922B |
| *Tetraphyllum roseum* Stapf | SAMN17002006 | P704 | Middleton, D. J. 5440 (E) [MMOG-281]; cult., RBGE, 20101826*A |
| *Titanotrichum oldhamii* (Hemsl.) Soler. | SAMN17002052 | P888 | Wang, B. 3525 (E); cult., RBGE, 19991767A |
| *Vanhouttea calcarata* Lem. | SAMN17002020 | P743 | Perret, M. 368 (G); cult., CJBG, 20111105NO |
| *Vanhouttea hilariana* Chautems | SAMN17002045 | P848 | Pires, S. et al. AC506 (CESJ); cult., CJBG, 20131994 N0 |
| **Outgroup** | | | |
| *Calceolaria tripartita* Ruiz & Pav. (Calceolariaceae) | SAMN17002011 | P897 | Zuluaga, A. 2893 (CUVC); wild, Colombia 2019 |
| *Cubitanthus alatus* (Cham. & Schltdl.) Barringer (Linderniaceae) | SAMN17002062 | P722 | Perret, M. 370 (G); cult., CJBG, 20111129JO |
| *Fraxinus excelsior* L. (Oleaceae) | SAMN17002030 | P757 | Perret, M. 371 (G); cult., CJBG, s.n. |

(*continued on next page*)

**Table 1** (*continued*)

| Species | Accession ID | Lab ID | Voucher specimen; provenance |
|---|---|---|---|
| *Jovellana sinclairii* (Hook.) Kraenzl. (Calceolariaceae) | SAMN17002046 | P882 | John Innes Hort. Inst. 1174 (E); cult., RBGE, 19,330,356 |
| *Olea europaea* L. (Oleaceae) | SAMN17002031 | P758 | Perret, M. 372 (G); cult., CJBG, s.n. |
| *Paulownia tomentosa* Steud. (Paulowniaceae) | SAMN17002029 | P756 | Perret, M. 373 (G); cult., CJBG, s.n. |
| *Peltanthera floribunda* Benth. (Peltantheraceae) | SAMN17002038 | P774 | Hammel 19,855 (MO); herbarium G |
| *Salvia pratensis* L. (Lamiaceae) | SAMN17002028 | P754 | Perret, M. 369 (G); wild, CJBG |

among related Lamiales lineages. For example, relationships among the major lineages of Gesneriaceae are still poorly resolved and there is still no firm consensus on the phylogenetic positions of taxa such as *Peltanthera*, *Sanango*, *Titanotrichum* and *Calceolariaceae* (Weber et al., 2013; APG IV, 2016). Recent phylogenomic approaches provide the opportunity to fill these gaps in the Gesneriaceae, but so far they have been applied in few groups to solve issues of incomplete lineage sorting and hybridization (in *Achimenes:* Roberts and Roalson, 2018; in *Cyrtandra:* Kleinkopf et al., 2019). In the present study, we developed a gene capture method for sequencing hundreds of nuclear genes simultaneously and evaluated the utility of this dataset for phylogenetic studies both at deep and shallow evolutionary levels within the Gesneriaceae.

Targeted sequencing has emerged as a standard phylogenomic method that outperforms Sanger sequencing approaches for addressing challenging problems in plant systematics (McKain et al., 2018). Compared to whole genome sequencing, gene capture is a reduced-representation method that targets a subset of the genome, thus decreasing the cost and the computational effort. The flexibility of probe design, and the number of targets makes gene capture a versatile method that has been shown to solve phylogenetic relationships at various taxonomical levels ranging from ancient radiations to recently diverged populations (Nicholls et al., 2015, de la Harpe et al., 2019). Lineage-specific bait kits including hundreds of single-copy nuclear loci with orthologs across a taxonomic group of interest have been developed for several groups of plants (Mandel et al., 2014; Heyduk et al., 2016; Mitchell et al., 2017; Herrando-Moraira et al., 2018; Couvreur et al., 2019; Kleinkopf et al., 2019; Loiseau et al., 2019; Soto Gomez et al., 2019). Complementary to these lineage-specific solutions, universal bait kits have been designed for applications across a wide breadth of angiosperm diversity (Buddenhagen et al., 2016; Léveillé-Bourret et al., 2018; Johnson et al., 2018). Recent results also demonstrated the utility of these universal kits to solve recent radiations when used alone or in combination with additional taxon-specific loci (Kriebel et al., 2019; Larridon et al., 2019; Murphy et al., 2020).

Here, we developed and tested the first sequence capture kit to perform phylogenetic analyses across the entire family Gesneriaceae. We strategically selected single-copy genes among the orthologous genes identified in the transcriptomic data available for Gesneriaceae (Chiara et al., 2013; Serrano-Serrano et al., 2017, 2019; Möller et al., pers. comm.) and designed custom baits in order to capture and sequence these selected genes across the family. Specifically, we aimed to 1) assess the performance of our bait kit and propose a selection of the most useful genomic regions to resolve phylogenetic relationships at different taxonomical levels across the Gesneriaceae and beyond; 2) compare the phylogenetic informativeness between the genes derived from the Universal Angiosperms-353 kit developed by Johnson et al. (2018) and our designed set of Gesneriaceae-specific genes; and 3) reconstruct a family-wide phylogeny for the Gesneriaceae using maximum likelihood and coalescent-based methods. Our results show that the molecular tools we developed here successfully generate highly-supported phylogenies for Gesneriaceae, offering new research opportunities to test hypotheses about the factors underlying the speciation and morphological diversification in the Gesneriaceae.

## 2. Materials and methods

### 2.1. Taxon sampling

Plant materials were collected from the living Gesneriaceae collections at the Conservatory and Botanical Garden of Geneva and the Royal Botanic Garden Edinburgh or collected in the wild and dried in silica gel (Table 1). Since the main aim of the sampling strategy was to assess the usefulness of our baits set across the Gesneriaceae, we included 70 Gesneriaceae samples from 52 genera representing 20 out of the 24 recognized subtribes in the family (Weber et al., 2013). We also selected 8 outgroup samples representing other Lamiales lineages including the monotypic genus *Peltanthera* and the family Calceolariaceae (*Calceolaria* and *Jovellana*), which have been identified as the closest relatives of the Gesneriaceae family (Perret et al., 2013; Refulio-Rodriguez and Olmstead, 2014; Angiosperm Phylogeny Group, 2016; Luna et al., 2019), and four other families within the order Lamiales (Lamiaceae, Linderniaceae, Oleaceae, and Paulowniaceae).

### 2.2. Target selection and bait design

We developed a Gesneriaceae bait set focusing on a wide range of target genes (Supplementary Table 1). In order to obtain a set of genes suitable for phylogenetic analyses, we first retained 7287 one-to-one orthologous groups (OG) present in the *de novo* transcriptome assemblies of six species from the New World subfamily Gesnerioideae (called New World hereafter): *Nematanthus albus*, *Nematanthus fritschii*, *Sinningia eumorpha*, *Sinningia magnifica*, *Paliavana tenuiflora* and *Vanhouttea calcarata* (Serrano-Serrano et al., 2017, 2019). Each of these OG were searched with BLAST (Altschul et al., 1990) in the transcriptomes of four species from the Old World Didymocarpoideae subfamily (called Old World hereafter); *Henckelia anachoreta*, *Leptoboea multiflora, Streptocarpus rexii* and *Streptocarpus glandulosissimus* (Möller et al., pers. comm.; Chiara et al., 2013) as well as in the genome of *Erythranthe guttata* (Mimulus Genome Project, DoE Joint Genome Institute; Nordberg et al., 2014). The Old World and New World Gesneriaceae sequences were combined and aligned using MAFFT v7.450 (Katoh and Standley, 2013). For each gene, we reported the number of corresponding hits found per species, and pairwise identities were calculated with Geneious v9.1.5 (https://www.geneious.com; Kearse et al., 2012). We selected 603 genes that were present in all 10 Gesneriaceae transcriptomes, and only have one BLAST hit per species. After removing the sequences that i) did not have a corresponding sequence in *E. guttata*; and ii) matched any mitochondrial or plastid genes of *Dorcoceras hygrometricum* (previously *Boea hygrometrica*; Xiao et al., 2015), we retained 551 genes for our bait set.

In addition to the gene set described above, we also included in our bait set a selection of the 353 genes suggested in a Universal Angiosperm bait kit (Johnson et al., 2018). We identified Gesneriaceae homologs corresponding to the Angiosperm-353 loci to use as reference for the Old World and New World Gesneriaceae. We removed any gene that gave BLAST hits of less than 70% percentage identity. In case a match was found in more than one reference, we kept the longest sequence. At the end, we retained 279 unique genes from the Angiosperms-353 probe set and added them to our initial gene selection. This expanded our final bait set to a total of 830 loci and a total length of 776,754 bp.

To allow uniform sequence recovery in both major clades of Gesneriaceae, two target sequences were designed per gene; one for the New World Gesneriaceae and one for the Old World Gesneriaceae. When the sequences were present in more than one species, we retained the longest one. For all targeted sequences, 80 bp long baits were designed and manufactured by Arbor Biosciences (Ann Arbor, MI). The Gesneriaceae targeted sequencing kit is publicly available at Zenodo under the name Gesneriaceae_830 (DOI: https://doi.org/10.5281/zenodo.4436683).

### 2.3. DNA extraction and library preparation

The silica gel-dried leaf samples (25–65 mg per sample) were homogenized using a TissueLyser II (Qiagen, Venlo, the Netherlands), and genomic DNA was extracted using a modified CTAB method (Doyle and Doyle, 1987; see Supplementary Methods for detailed protocol). DNA was quantified on a Qubit 3.0 fluorometer (Invitrogen, Thermo Fisher Scientific, Waltham, MA) using a High-sensitivity dsDNA Quantitation kit (Allsheng, Hangzhou, China) and visualized on a 2 % agarose gel.

In order to prepare libraries for each sample, 2,000 ng of genomic DNA in 100 μl $ddH_2O$ was sonicated using a Qsonica Q800R3 Sonicator (Qsonica, Newtown, CT). For each sample, 75 s of sonication with 25 % intensity was performed at 4 °C. The fragment size distribution for each sample was checked using a BiOptic Qsep100 Bio-Fragment Analyzer using the standard S2 cartridge and visualized on the Q-analyzer software (BiOptic, New Taipei City, Taiwan).

For fragment size selection, we prepared Serapure magnetic beads (Faircloth and Glenn, 2011; Rohland and Reich, 2012) and used a magnetic bead : DNA ratio of 1.0 in order to select for a fragment size range of 500 bp to 1,000 bp. We used KAPA HyperPrep kit (Roche, Basel, Switzerland) for the library preparation (see Supplementary Methods for detailed protocol). At the end of the library preparation, we checked the DNA quantity on a Qubit 3.0 fluorometer (Invitrogen, Thermo Fisher Scientific, Waltham, MA) using a High-sensitivity dsDNA Quantitation kit (Allsheng, Hangzhou, China).

### 2.4. Hybridization capture and sequencing

We pooled a total of 78 samples into three groups with the final DNA amount of 300–2800 ng per pool. We vacuum dried the pooled samples using a Savant SPD111V SpeedVac (Thermo Fisher Scientific, Waltham, MA) with a non-heated setting, and reconstituted the dried samples with 7 μl $ddH_2O$. Pooling of the samples was carried out using unique dual-indexing with combinations of 60 sequencing primers (Illumina, San Diego, CA).

For the hybridization capture, we used the myBaits® protocol (Arbor Biosciences, Ann Arbor, MI) following the manufacturer's guidelines with some modifications: The hybridization reaction was performed at 65 °C for 20 h. The post-hybridization library amplification was performed using the 2 × KAPA HiFi HotStart ReadyMix and the 10 × Library Amplification Primer Mix provided with the KAPA HyperPrep kit (Roche, Basel, Switzerland). Amplification reactions were performed in duplicates for each pool. The annealing time was set to 45 s per cycle, and cycle times were set to 12. The amplification reaction was purified using Serapure magnetic beads (Faircloth and Glenn, 2011; Rohland and Reich, 2012) with a magnetic beads : DNA ratio of 1.2.

Prior to sequencing, the DNA of the pooled samples was quantified on a Qubit 3.0 fluorometer (Invitrogen, Thermo Fisher Scientific, Waltham, MA) using a High-sensitivity dsDNA Quantitation kit (Allsheng, Hangzhou, China). The fragment size distribution for each pooled sample was checked on a BiOptic Qsep100 Bio-Fragment Analyzer using the S2 cartridge and visualized on the Q-analyzer software (BiOptic,New Taipei City, Taiwan). The duplicate pools were combined together prior to sending them for sequencing. 2 × 300 paired-end sequencing was performed on an Illumina MiSeq system (Illumina, San Diego, CA) at the iGE3 Genomics Platform, University of Geneva (Geneva, Switzerland).

### 2.5. Quality control, trimming, and mapping

The raw sequencing data was quality checked using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC (Ewels et al., 2016). Trimmomatic v0.39 (Bolger et al., 2014) was used to remove Illumina adapters and to filter out low-quality reads. Quality trimming was performed i) using a sliding window of 4 nucleotides and cutting a read when the average quality was lower than 15; ii) trimming the leading and trailing bases with a quality value lower than 20; and iii) removing the reads that were less than 40 bases long.

We used HybPiper (Johnson et al., 2016) to assemble the trimmed paired-reads and to generate consensus sequences. The pipeline with default settings can be briefly described in three steps: i) trimmed paired-reads were mapped using BWA (Li and Durbin, 2009); ii) the mapped reads were assembled into contigs using SPAdes (Bankevich et al., 2012); and iii) the assembled contigs were aligned to the reference target sequences using Exonerate (Slater and Birney, 2005). As an additional step, we used the "intronerate" function to retrieve off-target sequences as well as exons (collectively called "supercontigs) in fasta format.

### 2.6. Phylogenetic analyses

The phylogenetic analyses were performed using a subset of the target genes: We used several selection criteria to retain the most informative genes. We removed any gene that i) had less than 75% average length coverage; ii) was present in less than 75% of the samples; and iii) received paralog warnings for more than five samples in HybPiper.

The remaining genes were aligned using MAFFT v7.450 (Katoh and Standley, 2013) and concatenated using AMAS (Borowiec, 2016) to generate a supermatrix for phylogenetic inference analyses. Maximum likelihood was implemented using RAxML v8.2.4 (Stamatakis, 2014) with a GTRGAMMA substitution model for each gene and rapid bootstrap analysis with 100 bootstrap replicates. In order to estimate the species tree from the set of gene trees, a coalescent approach was performed using ASTRAL v5.6.3 (Zhang et al., 2018). Gene trees were generated using RAxML with the settings described above, and ASTRAL was used to compute quartet scores, which measure the level of congruence among the gene trees. The quartet scores were incorporated into the species tree using a previously developed R script (https://github.com/sidonieB/scripts/blob/master/plot_Astral_trees.R). In order to further quantify phylogenetic congruence, quadripartition internode certainty scores (QP-IC) were calculated using the program QuartetScores (Zhou et al., 2020). Providing the species-tree as reference and gene trees as input, QuartetScores quantifies the certainty for each internode within the species-tree while correcting for incomplete taxon sampling in the gene trees. While higher QP-IC scores indicate higher certainty for the internal nodes across the gene trees, lower scores indicate incongruency. All reconstructed phylogenetic trees were visualized with FigTree 1.4.4 (Rambaut, 2014).

## 3. Results

### 3.1. Target capture sequencing

We recovered an average of 689,700 raw reads for the New World Clade, 479,500 for the Old World Clade, and 317,000 for the outgroup taxa (Table 2). After the first quality filtering, we retained an average of 91%, 89%, and 82% of the raw reads respectively. Raw reads for all accessions are available at the GenBank Sequence Read Archive (SRA) under BioProject ID PRJNA684442.

Our bait set targeted a total of 830 genes among which 551 were specific to Gesneriaceae and 279 corresponded to genes listed in the Angiosperms-353 bait kit designed by Johnson et al. (2018). Gene lengths ranged from 128 bp to 3,663 bp, with an average of 955 bp. The

**Table 2**
Average summary statistics for the sequence capture.

| | # Reads (×1000) | Post-QC Survival % | % Genes with Sequences | Recovered Gene Length % |
|---|---|---|---|---|
| **GLOBAL** | 557.6 | 89.0 | 87.8 | 89.0 |
| **GESNERIOIDEAE** | 689.7 | 91.0 | 92.6 | 94.6 |
| **DIDYMOCARPOIDEAE** | 479.5 | 89.0 | 90.0 | 89.1 |
| **OUTGROUP** | 317.0 | 82.0 | 61.6 | 66.8 |

total sequence length of the 830 genes was 776,754 bp. The average gene recovery success was 87.8%, and the highest percentages of genes with sequences were observed within the New World clade (92.6%), followed by the Old World clade (90.0%), whereas the outgroup success was lower (61.6%; Table 2; Fig. 1). We observed a similar trend in recovered gene length, which was 94.6% in the New World clade, 89.1% in the Old World Clade, and 66.8% in the outgroup taxa. The outgroup performance was not correlated with phylogenetic distance (Fig. 2). For example, *Calceolaria* and *Jovellana*, the sister taxa of Gesneriaceae, had an average of 15.5% gene length coverage, whereas *Fraxinus*, one of the most distant outgroup taxa, had an average of 52.1% gene length coverage.

### 3.2. Extended phylogenetic dataset from non-targeted sequences

In addition to the targeted regions, we recovered long stretches of non-targeted sequences (Fig. 3). These sequences mostly include introns, but also stretches of downstream and upsteam regions of the targeted

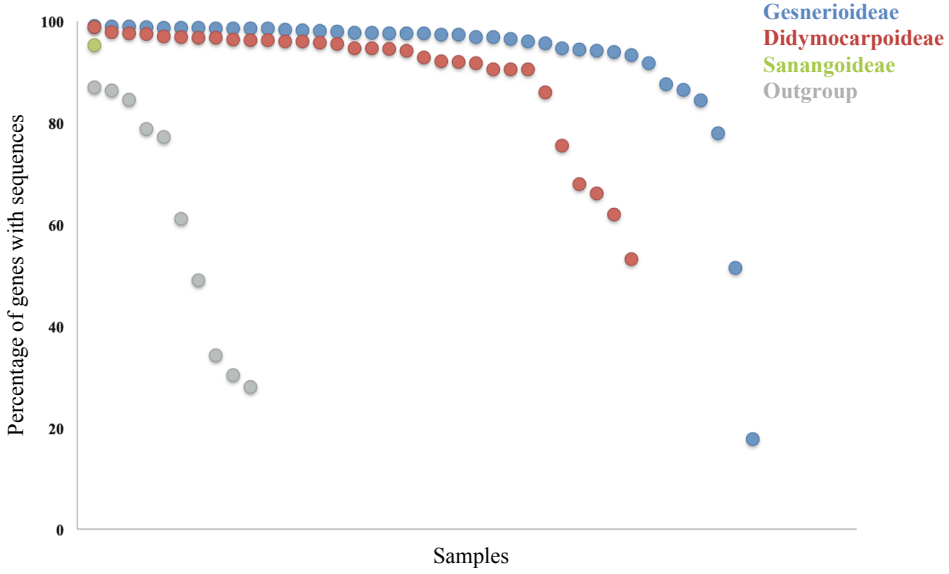


**Fig. 1.** Percentage of genes with sequences in our final set of 418 genes across all Gesneriaceae subfamilies. Each data point represents a sample

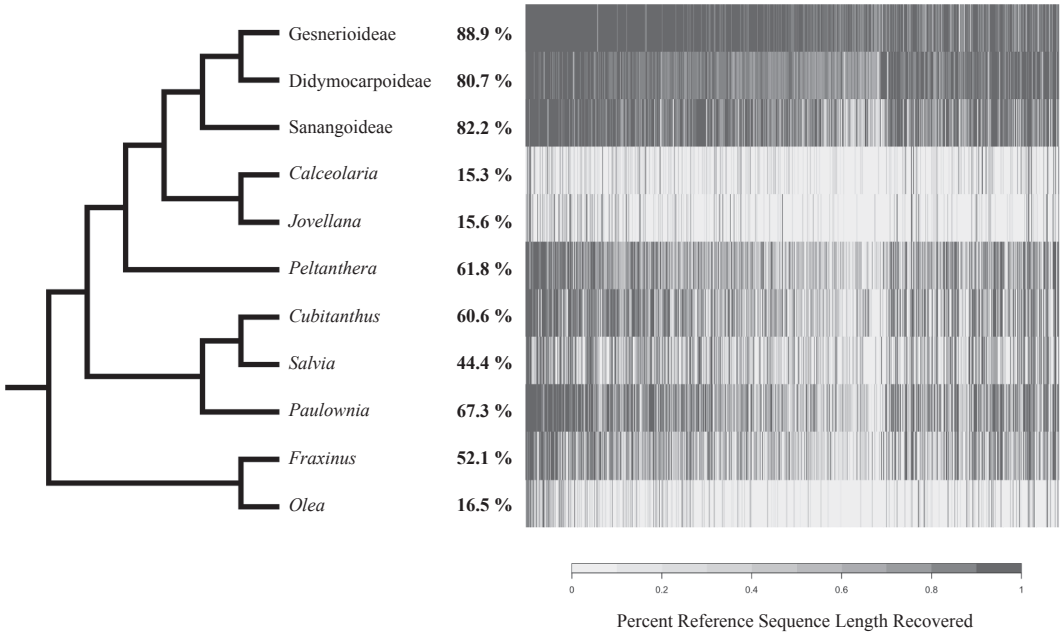


**Fig. 2.** Recovered sequence length heatmap for our final set of 418 genes. Each row corresponds to a taxonomic group, and each column corresponds to a gene. The shading of the bars represent the length of the recovered sequence relative to the reference gene. The percentage values for each taxonomic group represent the average sequence length recovered for the whole gene set.

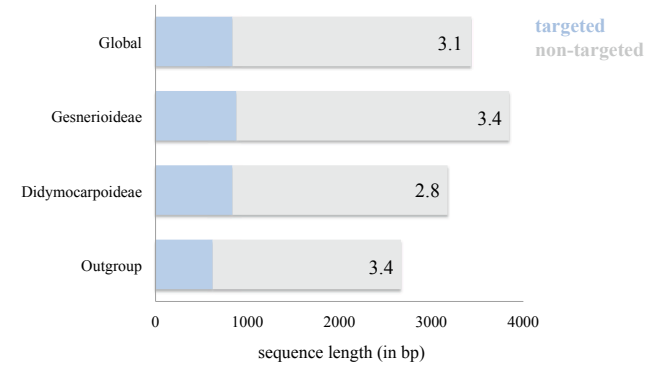


**Fig. 3.** Comparison of targeted versus non-targeted DNA sequences recovered for all samples in our final set of 418 genes. Numbers show the ratio of non-targeted to targeted sequence lengths.

**Table 3**
Alignment summary for the targeted regions versus supercontigs in the subtribes Streptocarpinae (Old World Didymocaropideae) and Ligeriinae (New World Gesnerioideae). × diff: fold difference between the target and supercontig values.

| | Streptocarpinae | | | Ligeriinae | | |
|---|---|---|---|---|---|---|
| | target | supercontig | x diff | target | supercontig | x diff |
| **Alignment Length** | 1044 | 4870 | 4.7 | 1110 | 6929 | 6.2 |
| **# Variable Sites** | 70 | 611 | 8.7 | 116 | 1283 | 11.1 |
| **% Variable Sites** | 6.8 | 12.8 | 1.9 | 10.3 | 19.0 | 1.8 |
| **# Parsimony Informative Sites** | 8 | 59 | 7.0 | 41 | 421 | 10.3 |
| **% Parsimony Informative Sites** | 0.9 | 1.2 | 1.4 | 3.7 | 6.3 | 1.7 |

genes. The average length of these regions was 2,597 bp, which was approximately 3 times longer than the average captured target region (832 bp). The captured off-target to target ratio was similar among all the groups, ranging from 2.8 within the Old World Clade to 3.4 within the New World Clade and the outgroup taxa.

The combination of targeted and non-targeted regions (called supercontigs hereafter) provided an extended dataset for closely related infrageneric taxa. When compared to the targeted regions, supercontigs provided 7.0 and 10.3 times more parsimony informative sites in Ligeriinae (*Paliavana, Sinningia*, and *Vanhouttea*) and Streptocarpinae (*Streptocarpus*), respectively (Table 3).

### 3.3. Genes selected for phylogenetic inference

Out of the initially targeted 830 genes, we retained 737 genes that were sequenced for ≥ 75% of their average length and for ≥ 75% of the samples. Among these genes, 219 were identified as probable paralogs according to HybPiper. The exclusion of these genes resulted in a final count of 418 genes suitable for phylogenetic inferences.

The 418 genes selected for phylogenetic applications had an average of 89.3% sample coverage (Table 4). The average length of the aligned genes was 1,223 bp, ranging from 284 bp to 3,245 bp. The size distribution of these genes differed between the two sets of loci: The gene length for the Gesneriaceae-specific loci had higher average values (1003 bp) than the Angiosperms-353 loci, which had an average of 825 bp. The percentage of variable sites and parsimony informative sites ranged from 37.0% to 70.8% and 23.3% to 54.8%, respectively. The Gesneriaceae-specific loci had a larger number of parsimony informative sites per locus than the Angiosperms-353 loci, which is mainly explained by their larger lengths, since the rate of parsimony informative characters per gene was similar between the two sets of loci (Fig. 4).

**Table 4**
Alignment summary for the 418 genes used in the phylogenetic reconstruction.

| | Average | Min | Max |
|---|---|---|---|
| **# Taxa** | 73 | 62 | 82 |
| **% Taxa** | 89.3 | 75.6 | 100.0 |
| **Alignment Length** | 1223 | 284 | 3245 |
| **% Variable Sites** | 56.7 | 37.0 | 70.8 |
| **% Parsimony Informative Sites** | 39.2 | 23.3 | 54.8 |

### 3.4. Family-wide phylogenetic reconstruction

The reconstructed ML tree derived from the analysis of the concatenated dataset of 418 genes (477,320 characters) was overall well resolved and most of the nodes were highly supported with bootstrap values of 95% or higher (Fig. 5). Calceolariaceae and *Peltanthera* were successively sister to the Gesneriaceae. Within Gesneriaceae, the monotypic genus *Sanango* representing the subfamily Sanangoideae was sister to the rest of the family. Monophyly of the subfamilies, Gesnerioideae and Didymocarpoideae, and of all the currently recognized tribes and subtribes was highly supported (bootstrap values = 100%). The species tree using a coalescent approach with the same 418 gene trees resulted in a topology identical to the ML tree, except for 3 nodes: the position of *Titanotrichum oldhamii*, *Streptocarpus formosus*, and the relationship between *Cyrtandra* and *Gyrocheilos* + *Didymocarpus*, which show a high level of gene tree incongruence as indicated by the quartet support values in the ASTRAL analysis (Fig. 6). The QP-IC scores ranged from 0.96 to 0.0 with an average of 0.36, meaning that there is high to moderate support for the reference topology throughout the phylogeny (Supplementary Figure S1). QP-IC scores were consistent with the quartet support values, both of which were lower at the three nodes mentioned above. The lack of negative scores throughout the phylogeny suggested that the topology of the species tree was more frequent than any of the conflicting alternative topologies of the individual gene trees.

## 4. Discussion

Resolving relationships with highly supported phylogenies is a prerequisite for many downstream applications such as investigating diversification rates, unravelling biogeographic history, and inferring the timing of evolutionary events. Previous phylogenetic studies on plant taxa relied on small sets of markers, which do not always have the power to resolve phylogenetic relationships due to their low numbers of informative characters (Parks et al., 2009; Fragoso-Martínez et al., 2017). Current practice in plant phylogenomics is to develop clade-specific bait kits designed to capture several hundred loci and utilize this large dataset to perform high-resolution phylogenetic reconstructions (e.g., Couvreur et al., 2019; Loiseau et al., 2019; Moore et al., 2018).

The Gesneriaceae have been the subject of several large-scale phylogenetic analyses although most approaches have deficiencies in providing highly supported and fully resolved trees. This might be due to the low number of phylogenetically informative characters and missing data (e.g. Wortley et al., 2005; Möller et al., 2009; Roalson and Roberts, 2016). As a consequence, some of the subtribal and tribal relationships in the family are not fully understood and several uncertainties exist (Möller and Clark, 2013). Here, we addressed these deficiencies by developing a method to simultaneously obtain molecular sequence data of hundreds of genes with a wide range of evolutionary rates making them applicable to a wide taxonomic range.

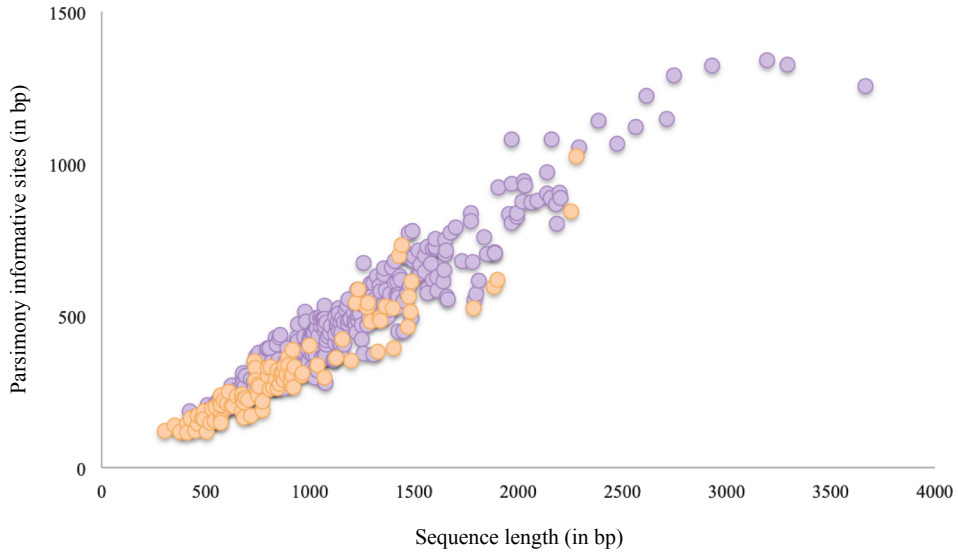


**Fig. 4.** Sequence length versus parsimony informative sites for the Gesneriaceae-specific genes (purple) and for the genes from the Angiosperms-353 bait set (orange). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

### *4.1. A bait kit for phylogenetic inference across Gesneriaceae and beyond*

Our Gesneriaceae bait kit enabled the sequencing of 830 loci, representing an aligned total of 776,754 bp. The high recovery success of the targeted regions in all Gesneriaceae subfamilies (92.6% in Gesnerioideae; 90.0% in Didymocarpoideae; 95.3% in Sanangoideae; Fig. 1; Table 2) demonstrated the high bait efficiency across the entire family. These homogeneous results across the family may be due to the use of several reference transcriptomes for bait design and the use of baits that include variants of each gene from the New World and the Old World Gesneriaceae. A recently developed probe set targeting *Cyrtandra* includes 570 loci with an average target length of 317 bp and 12.6% parsimony informative characters (Kleinkopf et al., 2019). This genus-specific bait set showed no overlap with our bait set except for a single target that was present in both sets (OG7527; See Supplementary Table S1 for locus information). Overall, our family-wide bait set included more loci (830) with larger average length (955 bp) and higher percentage of parsimony informative characters (39.2%), and was aimed to be utilized at a broader taxonomic range, from species to family level. For the outgroup samples, the gene recovery rates of our bait kit ranged from 65% to 15% (Fig. 2), which indicates that the bait kit could be used at a larger taxonomical level to resolve phylogenetic questions across the Lamiales, although the recovery success of the bait kit outside the Gesneriaceae remains to be further evaluated using a broader taxon sampling.

To expand the phylogenetic application of our bait kit, we supplemented our Gesneriaceae-specific loci with the genes from the Angiosperms-353 probe set using Gesneriaceae-specific *de novo* baits. This approach enabled us to successfully capture the Angiosperms-353 genes, while improving the specificity. We show that the Gesneriaceae-specific loci provided approximately five times more parsimony informative characters than the loci from the Angiosperms-353 kit. However, the ratio of parsimony informative characters to gene length was comparable between the Angiosperms-353 and the Gesneriaceae-specific loci (Fig. 4), in agreement with an earlier demonstration that family-specific kits in plants do not necessarily have more phylogenetic power than the universal kits (Larridon et al., 2019).

### *4.2. Next-generation phylogeny of Gesneriaceae*

Here, we present the first phylogenetic reconstruction across the family Gesneriaceae using targeted gene capture. After excluding paralogs and genes with length and sample coverage lower than 75%, we retained a subset of 418 genes suitable for the phylogenetic analyses of our 70 samples representing all tribes and 20 out of 24 subtribes recognized in the family (Table 1). Our phylogenies reconstructed using concatenation-based (Fig. 5) and coalescent-based (Fig. 6) approaches were greatly congruent among deep level relationships, except for the position of *Titanotrichum* which was sister to tribe Beslerieae in the former, and sister to the clade Coronanthereae + Gesnerieae in the latter. This conflicting placement of *Titanotrichum* correlates with a high level of gene tree incongruence as revealed by the quartet support and QP-IC values (Fig. 6, Supplementary Figure S1). This observation, coupled with the short branch lengths separating *Titanotrichum* from its closest relative Napeantheae and Beslerieae (Fig. 5), points to the possibility of incomplete lineage sorting following rapid divergence as a likely explanation for these gene tree discordances and the still contentious phylogenetic position of this Asiatic genus within the New World Gesneriaceae (Wang et al., 2004; Möller and Clark, 2013; Roalson and Roberts, 2016). It is interesting to note that the placement of *Titanotrichum* in the concatenated tree as sister to Beslerieae is identical in the comprehensively sampled analysis of Luna et al. (2019) who used four chloroplast marker sequences. Such confirmation from a chloroplast dataset of the results of the nuclear analysis here is strong evidence for this relationship.

Overall our results agree with the latest formal classification of Gesneriaceae (Weber et al., 2013) and the latest phylogenetic analyses performed at the family-wide scale (Roalson and Roberts, 2016; Luna et al., 2019). Our phylogenetic trees recovered all three subfamilies as monophyletic, with the monotypic Sanangoideae as sister to the rest. The monophyly of all tribes and subtribes (where more than one sample was included) was also recovered with high support and in agreement with Weber et al. (2013).

Our results also provide insights into the phylogeny of Gesneriaceae and the placement of this family within Lamiales. Calceolariaceae is here identified as the sister family of Gesneriaceae with the monotypic genus *Peltanthera* sister to both families. This result is in agreement with an angiosperm-wide analysis of Soltis et al. (2011) and an extensive analysis of Gesneriaceae and Lamiales (Luna et al., 2019), but conflicts with other Gesneriaceae-focused studies that placed this taxon sister to the Gesneriaceae family (Perret et al., 2013, Roalson and Roberts, 2016).

In the Gesnerioideae, our results support the position of tribe Napeantheae as the first diverging lineage in the Gesnerioideae. Previous analyses including this clade were either congruent with our result

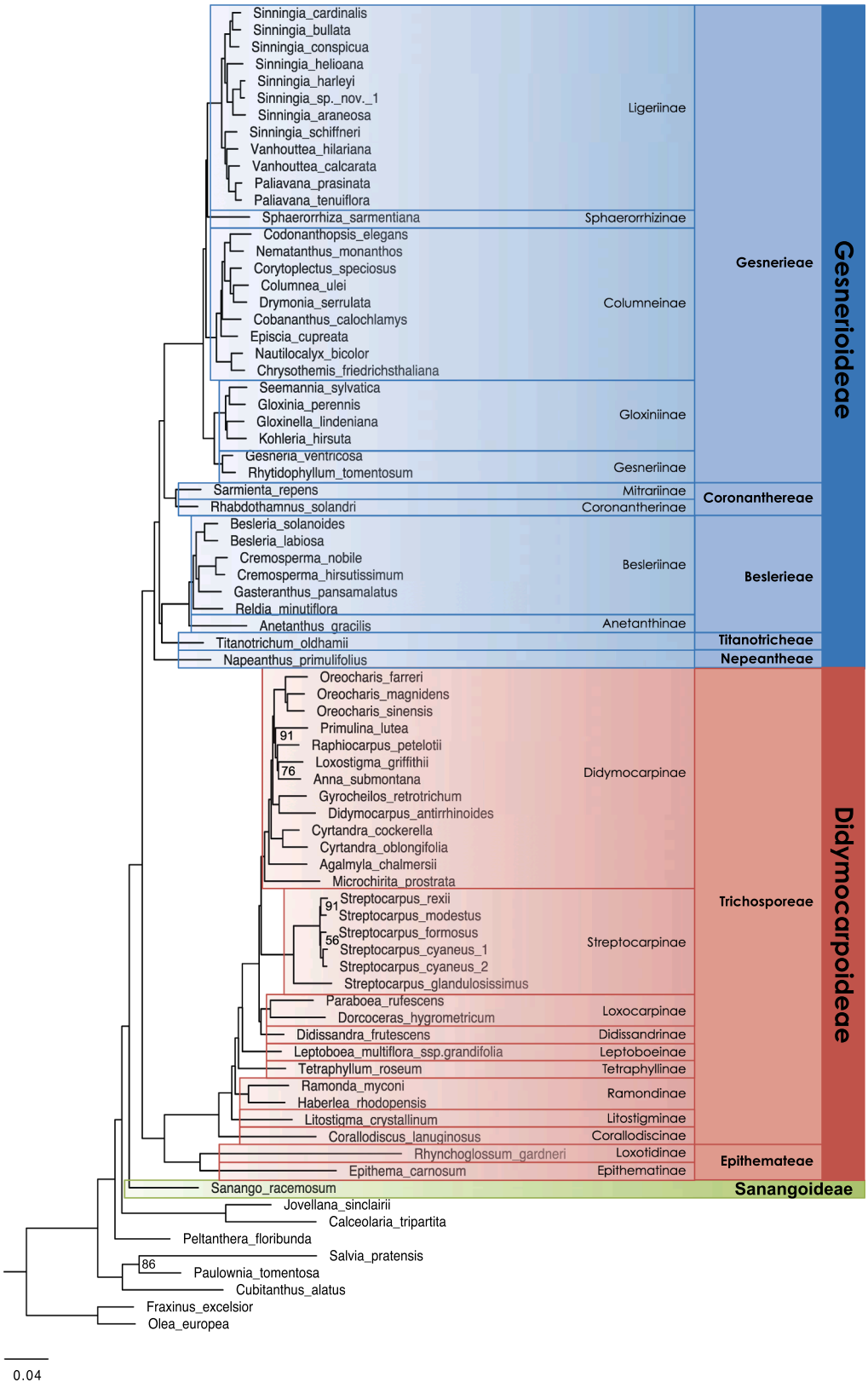


**Fig. 5.** Maximum-likelihood-based phylogenetic reconstruction of the Gesneriaceae generated using a supermatrix of 418 gene sequences for 78 samples. Values on the nodes represent bootstrap values. Bootstrap values above 95% are not shown. Subribes (left) and tribes (middle) of th subfamilies (right) Gesnerioideae (blue), Didymocarpoideae (red) and Sanangoideae (green) are highlighted in boxes. Classification of Gesneriaceae follows Weber et al. (2013). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(Roalson and Roberts, 2016; Luna et al., 2019) or recovered Napeantheae as a sister to Beslerieae with low support (Perret et al., 2013). The Sphaerorrhizinae is recovered as the sister clade of Ligeriinae, which is in agreement with the overlapping geographical distribution of these subtribes in Brazil (Perret et al., 2013; Araujo et al., 2016). Previously, this species-poor clade has been variously related to other subtribes within the Gesnerieae with low support (Zimmer et al., 2002; Araujo et al., 2016; Roalson and Roberts, 2016).

In the Didymocarpinae, the position of the *Didymocarpus/Gyrocheilos* clade varied between the two tree building methods: in the concatenated

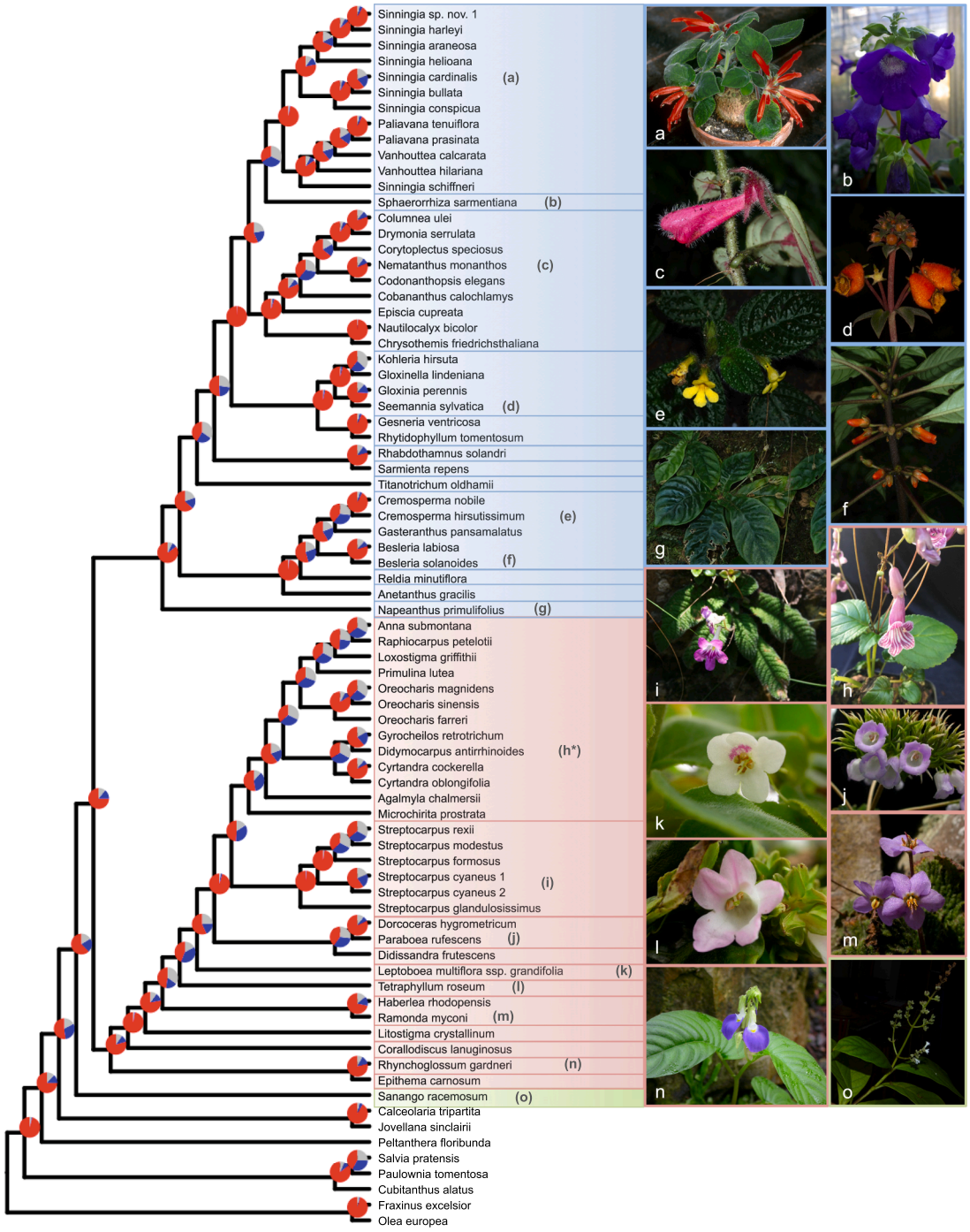


**Fig. 6.** Coalescent-based phylogenetic reconstruction of the Gesneriaceae inferred using the set of 418 gene trees. Pie charts on the nodes represent the percentage of the gene trees agreeing with the topology of the main species tree (red) and the other two alternative topologies (blue and gray). Gesnerioideae (blue) and Didymocarpoideae (red) subtribes are highlighted in boxes as in the Fig. 5. Photos by Alain Chautems (a), Mathieu Perret (b, c, e, f, g), John L. Clark (d, o), Franz Xaver (i), and Michael Möller (h, j, k, l, m, n). The photo (h) is *Didymocarpus purpureobracteatus* (instead of *D. anthirrhinoides*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

analysis it was sister to *Oreocharis*, while in the coalescent-based approach it was sister to *Cyrtandra*. In Roalson and Roberts (2016), *Oreocharis* and *Didymocarpus/Gyrocheilos* were sister clades, which might indicate that the concatenation-based approach might be a better reflection of this relationship. This incongruence may be due to the limited sampling in this subtribe and the very short backbone branches in Didymocarpinae. This is similar to previous analyses of other markers such as ITS and *trn*LF (Möller et al., 2011) and might suggest the presence of a radiation in the diversification history of the subtribe.

In tribe Trichosporeae, the relationships of several subtribes were clarified compared to the previously published trees (Möller and Clark, 2013; Weber et al., 2013). Considering the absence of subtribe Jerdoniinae, the first branch to split off was occupied by Corallodiscinae as previously reported, but on the following grades were Litostigminae, Ramondinae, and then Tetraphyllinae and Leptoboeinae. Tetraphyllinae was regarded as an earlier, though unsupported, divergent lineage in previous analyses (e.g. Möller et al., 2009). The remaining four subtribes formed two sister pairs: Didissandrinae and Loxocarpinae as one; and

Streptocarpinae and Didymocarpinae as the other. This topology is discordant with the previous ones (Möller and Clark, 2013; Roalson and Roberts, 2016), and this might be due to low sampling in the present study. However, at species level, comparisons within *Streptocarpus* revealed that while the species of the Cape Primrose clade formed a polytomy in previous studies using the nuclear ITS alone or incombination with chloroplast *rpl*20-*rps*12, and *trn*LF sequence data (Möller and Cronk 2001a,b; Nishii et al., 2015), here they had fully resolved and highly supported relationships. This demonstrated the power of the gene capture approach over conventional Sanger sequencing-based approaches.

### 4.3. Potential use of paralogs and non-targeted regions

Out of the initial 830 sequenced loci, 219 were marked by HybPiper as potential paralogs, and we excluded them from our phylogenetic analyses for which we had sufficient data to obtain high resolutions and topology support. This was a conservative decision, which removed a substantial amount of sequence data that could potentially have been used in many downstream analyses. While earlier phylogenetic pipelines utilized only single-copy genes, there are now several methods to incorporate paralogs to enrich phylogenetic analyses (Yang and Smith, 2014; Moore et al., 2018 Koenen et al., 2020; Zhang et al. 2020). In our case, inclusion of these potentially paralogous loci would increase the sequence data by more than 25%, which could be invaluable in resolving species-level phylogenies or studying population structures in the future.

Outside the targeted regions, we also covered a significant portion of the non-targeted regions, including introns and intergenic sequences. When compared to the coding regions of the genome, introns and intergenic sequences have faster mutation rates and more neutral evolution and are therefore considered as valuable phylogenetic markers (Creer, 2007; Irimia and Roy, 2008). The captured length of these non-targeted regions in the present study was approximately 3 times longer than the targeted regions (Fig. 3), and they can substantially increase the genetic information to be used in downstream phylogenetic analyses. In Gesneriaceae, some taxa went through rapid and recent radiations, and the phylogenetic structures within these clades are still unclear. Dispersal to new environments and adaptations to different habitats contributed to the high rate of diversification events in the neotropical Ligeriinae and the paleotropical Streptocarpinae, (Möller and Cronk, 2001b; Perret et al., 2007; Roalson and Roberts, 2016). In these highly diverse subtribes, the number of phylogenetically informative characters increased up to ten fold when we supplemented the targeted sequences with non-targeted regions (Table 3). This demonstrates the great potential of this supercontig dataset that can be applied in the future to resolve phylogenetic relationships within and among other genera that have been difficult to study with standard genetic markers.

## 5. Conclusion

Here we outlined our approach in designing baits to generate nuclear DNA sequence data useful for family wide and species level phylogenetic analyses in the Gesneriaceae. Our bait set enabled the capture of 830 genes, among which 551 were specific to Gesneriaceae and 279 were from the Angiosperms-353 baiting kit designed by Johnson et al. (2018). We captured these 830 genes across the Gesneriaceae with a high recovery success and showed the potential applicability of our bait-kit in other Lamiales families. After screening for non-paralogs and phylogenetic informativeness, we retained 418 loci, which provided sufficient phylogenetic signal to resolve relationships from species to family level, confirming previously indicated relationships and providing additional resolution on previously intractable relationships. Our strategy of combining taxon-specific and more universal sets of loci in a single baiting kit has clear advantages: while the angiosperm universal loci allow data reuse to contribute to the efforts towards the assemblage of the plant Tree of Life (Eiserhardt et al., 2018), the family-specific loci will provide added support and resolution to the Gesneriaceae phylogeny and new opportunities to explore diversification of this plant lineage at different taxonomic levels.

## CRediT authorship contribution statement

**Ezgi Ogutcen:** Conceptualization, Methodology, Software, Investigation. **Camille Christe:** Methodology, Software. **Kanae Nishii:** Resources, Methodology. **Nicolas Salamin:** Methodology, Software. **Michael Möller:** Resources, Conceptualization. **Mathieu Perret:** Conceptualization, Resources, Supervision.

## Declaration of competing interest

The authors declare that they have no competing interests that could have influenced the work reported in this paper.

*Acknowledgements*

We would like to thank Régine Niba, Antonio Castro Bareiro, and Dani Cardoso for their contribution to the laboratory work; Alain Chautems, Yvonne Menneret and the greenhouse staff in charge of the propagation and maintenance of the Gesneriaceae collection at the Botanical Garden of Geneva. We would also like to thank the Science and Horticultural divisions of the Royal Botanic Garden Edinburgh (RBGE) for their support. RBGE is supported by the Rural and Environment Science and Analytical Services Division (RESAS) in the Scottish Government. We thank Alain Chautems, Mauro Peixoto, Francisco Tobar, Laura Clavijo and Alejandro Zuluaga for their help in obtaining material used in this study. Collection permits were granted by the CNPq in Brazil (CMC 038/03) and the Autoridad Nacional de Licencias Ambientales (ANLA) from the Republica de Colombia no.1070 and resolución nos. 1004 and 1255. Images in figure 6 were kindly provided by Alain Chautems, John L. Clark and Franz Xaver under the CC-by-SA license. MP is funded by the Ville de Genève, and the Swiss National Science Foundation (Grant number: 31003A_175655). EO is financially supported by the Fondation Ernst et Lucie Schmidheiny. KN is financially supported by the Edinburgh Botanic Garden (Sibbald) Trust, Japan Society for the Promotion of Science (JSPS KAKENHI; Grant Number: 18K06375), and the Sumitomo Foundation.

*Data availability*

The following files are deposited at Zenodo and they are accessible for free (DOI: 10.5281/zenodo.4436683): Gesneriaceae baits, the reference target sequences for the baits, the list of target genes, coalescent-based species tree, concatenation-based supermatrix tree, individual gene trees.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2021.107068.

## References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Angiosperm Phylogeny Group, 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot. J. Linn. Soc. 181, 1–20.

Araujo, A.O., Chautems, A., Cardoso-Gustavson, P., Souza, V.C., Perret, M., 2016. Taxonomic revision and phylogenetic position of the Brazilian endemic genus *Sphaerorrhiza* (Sphaerorrhizineae, Gesneriaceae) including two new species. Syst. Bot. 41, 651–664.

Atkins, H.J., Bramley, G.L., Johnson, M.A., Kartonegoro, A., Nishii, K., Kokubugata, G., Moeller, M., Hughes, M., 2019. A molecular phylogeny of Southeast Asian *Cyrtandra*

(Gesneriaceae) supports an emerging paradigm for Malesian plant biogeography. Front. Biogeogr. e44814.
Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., et al., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.
Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.
Borowiec, M.L., 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4, e1660.
Buddenhagen, C., Lemmon, A.R., Lemmon, E.M., Bruhl, J., Cappa, J., Clement, W.L., et al., 2016. Anchored phylogenomics of angiosperms I: Assessing the robustness of phylogenetic estimates. bioRxiv 086298.
Chiara, M., Horner, D.S., Spada, A., 2013. De novo assembly of the transcriptome of the non-model plant *Streptocarpus rexii* employing a novel heuristic to recover locus-specific transcript clusters. PLoS One 8 (12).
Clark, J.L., Funke, M.M., Duffy, A.M., Smith, J.F., 2012. Phylogeny of a Neotropical clade in the Gesneriaceae: more tales of convergent evolution. Int. J. Plant Sci. 173, 894–916.
Clark, J.L., Herendeen, P.S., Skog, L.E., Zimmer, E.A., 2006. Phylogenetic relationships and generic boundaries in the Episcieae (Gesneriaceae) inferred from nuclear, chloroplast, and morphological data. Taxon 55 (2), 313–336.
Couvreur, T.L.P., Helmstetter, A.J., Koenen, E.J.M., Bethune, K., Brandão, R.D., Little, S. A., et al., 2019. Phylogenomics of the major tropical plant family Annonaceae using targeted enrichment of nuclear genes. Front. Plant Sci. 9, 1941.
Creer, S., 2007. Choosing and using introns in molecular phylogenetics. Evol Bioinform. 3, 117693430700300011.
de La Harpe, M., Hess, J., Loiseau, O., Salamin, N., Lexer, C., Paris, M., 2019. A dedicated target capture approach reveals variable genetic markers across micro-and macro-evolutionary time scales in palms. Mol. Ecol. Resour. 19, 221–234.
Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19, 11–15.
Eiserhardt, W.L., Antonelli, A., Bennett, D.J., Botigué, L.R., Burleigh, J.G., Dodsworth, S., et al., 2018. A roadmap for global synthesis of the plant tree of life. Am. J. Bot. 105 (3), 614–622.
Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32 (19), 3047–3048.
Faircloth, B., Glenn, T., 2011. Serapure protocol. V2.2. Ecol. and Evol. Biology.
Fragoso-Martínez, I., Salazar, G.A., Martínez-Gordillo, M., Magallón, S., Sánchez-Reyes, L., Moriarty Lemmon, E., et al., 2017. A pilot study applying the plant Anchored Hybrid Enrichment method to New World sages (*Salvia* subgenus Calosphace; Lamiaceae). Mol. Phylogenet. Evol. 117, 124–134.
Herrando-Moraira, S., Calleja, J.A., Carnicero, P., Fujikawa, K., Galbany-Casals, M., Garcia-Jacas, N., Im, H.T., Kim, S.C., Liu, J.Q., López-Alvarado, J., López-Pujol, J., 2018. Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe *Cardueae* (Compositae). Mol. Phylogenet Evol. 128, 69–87.
Heyduk, K., Trapnell, D.W., Barrett, C.F., Leebens-Mack, J., 2016. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. Biol. J. Linn. Soc. 117, 106–120.
Irimia, M., Roy, S.W., 2008. Spliceosomal introns as tools for genomic and evolutionary analysis. Nucleic Acids Res. 36 (5), 1703–1712.
Johnson, M.G., Gardner, E.M., Liu, Y., Medina, R., Goffinet, B., Shaw, A.J., Zerega, N.J. C., Wickett, N.J., 2016. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl. Plant Sci. 4, 1600016.
Johnson, M.G., Pokorny, L., Dodsworth, S., Botigué, L.R., Cowan, R.S., Devault, A., et al., 2018. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. Syst. Biol. 68 (4), 594–606.
Jong, K., Burtt, B.L., 1975. The evolution of morphological novelty exemplified in the growth patterns of some Gesneriaceae. New Phytol. 75 (2), 297–311.
Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780.
Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28 (12), 1647–1649.
Kleinkopf, J.A., Roberts, W.R., Wagner, W.L., Roalson, E.H., 2019. Diversification of Hawaiian *Cyrtandra* (Gesneriaceae) under the influence of incomplete lineage sorting and hybridization. J. Syst. Evol. 57 (6), 561–578.
Koenen, E.J., Ojeda, D.I., Bakker, F.T., Wieringa, J.J., Kidner, C., Hardy, O.J., Hughes, C. E., 2020. The Origin of the Legumes is a Complex Paleopolyploid Phylogenomic Tangle closely associated with the Cretaceous-Paleogene (K-Pg) Mass Extinction Event. Syst. Biol. 1–19.
Kriebel, R., Drew, B.T., Drummond, C.P., González-Gallegos, J.G., Celep, F., Mahdjoub, et al., 2019. Tracking temporal shifts in area, biomes, and pollinators in the radiation of *Salvia* (sages) across continents: leveraging anchored hybrid enrichment and targeted sequence data. Am. J. Bot. 106 (4), 573–597.
Larridon, I., Villaverde, T., Zuntini, A.R., Pokorny, L., Brewer, G.E., Epitawalage, N., et al., 2019. Tackling rapid radiations with targeted sequencing. Front. Plant Sci. 10, 1655.
Léveillé-Bourret, É., Starr, J.R., Ford, B.A., Lemmon, E.M., Lemmon, A.R., 2018. Resolving rapid radiations within angiosperm families using anchored phylogenomics. Syst. Biol. 67, 94–112.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.
Loiseau, O., Olivares, I., Paris, M., de La Harpe, M., Weigand, A., Koubinova, D., et al., 2019. Targeted capture of hundreds of nuclear genes unravels phylogenetic relationships of the diverse Neotropical palm tribe Geonomateae. Front. Plant Sci. 10, 864.
Luna, J.A., Richardson, J.E., Nishii, K., Clark, J.L., Möller, M., 2019. The family placement of *Cyrtandromoea*. Syst Bot. 44 (3), 616–630.
Mandel, J.R., Dikow, R.B., Funk, V.A., Masalia, R.R., Staton, S.E., Kozik, A., Michelmore, R.W., Rieseberg, L.H., Burke, J.M., 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. Appl. Plant Sci. 2 (2), 1300085.
McKain, M.R., Johnson, M.G., Uribe-Convers, S., Eaton, D., Yang, Y., 2018. Practical considerations for plant phylogenomics. Appl. Plant Sci. 6, e1038.
Mitchell, N., Lewis, P.O., Lemmon, E.M., Lemmon, A.R., Holsinger, K.E., 2017. Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *Protea* L. Am J Bot. 104 (1), 102–115.
Moore, A.J., Vos, J.M.D., Hancock, L.P., Goolsby, E., Edwards, E.J., 2018. Targeted enrichment of large gene families for phylogenetic inference: phylogeny and molecular evolution of photosynthesis genes in the Portullugo clade (Caryophyllales). Syst. Biol. 67 (3), 367–383.
Möller, M., Clark, J.L., 2013. The state of molecular studies in the family Gesneriaceae: a review. Selbyana 95–125.
Möller, M., Cronk, Q.C., 2001a. Evolution of morphological novelty: A phylogenetic analysis of growth patterns in *Streptocarpus* (Gesneriaceae). Evolution 55 (5), 918–929.
Möller, M., Cronk, Q.C., 2001b. Phylogenetic studies in *Streptocarpus* (Gesneriaceae): reconstruction of biogeographic history and distribution patterns. Syst. Geogr. Plants. 545–555.
Möller, M., Pfosser, M., Jang, C.G., Mayer, V., Clark, A., Hollingsworth, M.L., et al., 2009. A preliminary phylogeny of the 'didymocarpoid Gesneriaceae' based on three molecular data sets: Incongruence with available tribal classifications. Am. J. Bot. 96 (5), 989–1010.
Möller, M., Forrest, A., Wei, Y.G., Weber, A., 2011. A molecular phylogenetic assessment of the advanced Asiatic and Malesian didymocarpoid Gesneriaceae with focus on non-monophyletic and monotypic genera. Plant Syst. Evol. 292, 223–248.
Möller, M., Wei, Y., Wen, F., Clark, J.L., Weber, A., 2016. You win some you lose some: updated delineations and classification of Gesneriaceae – implications for the family in China. Guihaia 36 (1), 44–60.
Murphy, B., Forest, F., Barraclough, T., Rosindell, J., Bellot, S., Cowan, R., Golos, M., Jebb, M., Cheek, M., 2020. A phylogenomic analysis of *Nepenthes* (Nepenthaceae). Mol. Phylogenetics Evol. 144, 106668.
Nicholls, J.A., Pennington, R.T., Koenen, E.J., Hughes, C.E., Hearn, J., Bunnefeld, L., Dexter, K.G., Stone, G.N., Kidner, C.A., 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). Front. Plant Sci. 6, 710.
Nishii, K., Hughes, M., Briggs, M., Haston, E., Christie, F., DeVilliers, et al., 2015. *Streptocarpus* redefined to include all Afro-Malagasy Gesneriaceae: Molecular phylogenies prove congruent with geographical distribution and basic chromosome numbers and uncover remarkable morphological homoplasies. Taxon 64 (6), 1243–1274.
Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., et al., 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. 42, D26–D31.
Parks, M., Cronn, R., Liston, A., 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 7 (1), 84.
Perret, M., Chautems, A., Spichiger, R., Kite, G., Savolainen, V., 2003. Systematics and evolution of tribe Sinningieae (Gesneriaceae): evidence from phylogenetic analyses of six plastid DNA regions and nuclear ncpGS. Am. J. Bot. 90 (3), 445–460.
Perret, M., Chautems, A., Spichiger, R., Barraclough, T.G., Savolainen, V., 2007. The geographical pattern of speciation and floral diversification in the Neotropics: the tribe Sinningieae (Gesneriaceae) as a case study. Evolution 62, 1641–1660.
Perret, M., Chautems, A., De Araujo, A.O., Salamin, N., 2013. Temporal and spatial origin of Gesneriaceae in the New World inferred from plastid DNA sequences. Bot. J. Linn. Soc. 171 (1), 61–79.
Rambaut, A., 2014. FigTree 1.4. 2 software. Institute of Evolutionary Biology, Univ. Edinburgh.
Refulio-Rodriguez, N.F., Olmstead, R.G., 2014. Phylogeny of Lamiidae. Am. J. Bot. 101 (2), 287–299.
Roalson, E.H., Roberts, W.R., 2016. Distinct processes drive diversification in different clades of Gesneriaceae. Syst. Biol. 6, 662–684.
Roalson, E.H., Boggan, J.K. and Skog, L.E., 2005. Reorganization of tribal and generic boundaries in the Gloxinieae (Gesneriaceae: Gesnerioideae) and the description of a new tribe in the Gesnerioideae, Sphaerorrhizeae. Selbyana, 225–238.
Roberts, W.R., Roalson, E.H., 2018. Phylogenomic analyses reveal extensive gene flow within the magic flowers (*Achimenes*). Am. J. Bot. 105 (4), 726–740.
Rohland, N., Reich, D., 2012. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome res. 22 (5), 939–946.
Schulte, L.J., Clark, J.L., Novak, S.J., Ooi, M.T.Y., Smith, J.F., 2014. Paraphyly of section Stygnanthe (*Columnea*, Gesneriaceae) and a revision of the species of section Angustiflorae, a new section inferred from ITS and chloroplast DNA data. Syst. Bot. 39 (2), 613–636.
Serrano-Serrano, M.L., Rolland, J., Clark, J.L., Salamin, N., Perret, M., 2017. Hummingbird pollination and the diversification of angiosperms: an old and successful association in Gesneriaceae. Proc. R. Soc. B Biol. Sci. 284.
Serrano-Serrano, M.L., Marcionetti, A., Perret, M., Salamin, N., 2019. Convergent changes in gene expression associated with repeated transitions between hummingbird and bee pollinated flowers. bioRxiv 706127.

Slater, G.S.C., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6 (1), 31.
Soltis, D.E., Smith, S.A., Cellinese, N., Wurdack, K.J., Tank, D.C., Brockington, S.F., et al., 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am. J. Bot. 98 (4), 704–730.
Soto Gomez, M., Pokorny, L., Kantar, M.B., Forest, F., Leitch, I.J., Gravendeel, B., Wilkin, P., Graham, S.W., Viruel, J., 2019. A customized nuclear target enrichment approach for developing a phylogenomic baseline for *Dioscorea* yams (Dioscoreaceae). Appl. Plant Sci. 7 (6), e11254.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
Wang, C.N., Möller, M., Cronk, Q.C.B., 2004. Phylogenetic position of *Titanotrichum oldhamii* (Gesneriaceae) inferred from four different gene regions. Syst Bot. 29, 407–418.
Weber, A., Clark, J.L., Möller, M., 2013. A new formal classification of Gesneriaceae. Selbyana 31, 68–94.
Wortley, A.H., Rudall, P.J., Harris, D.J., Scotland, R.W., 2005. How much data are needed to resolve a difficult phylogeny? Case study in Lamiales. Syst Biol. 54, 697–709.
Xiao, L., Yang, G., Zhang, L., Yang, X., Zhao, S., Ji, Z., et al., 2015. The resurrection genome of *Boea hygrometrica*: A blueprint for survival of dehydration. Proc. Natl. Acad. Sci. USA 112 (18), 5833–5837.
Yang, Y., Smith, S.A., 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol. Biol. Evol. 31 (11), 3081–3092.
Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S., 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinf. 19, 153.
Zhang, C., Scornavacca, C., Molloy, E.K., Mirarab, S., 2020. ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. Mol. Biol. Evol. 37 (11), 3292–3307.
Zhou, X., Lutteropp, S., Czech, L., Stamatakis, A., Looz, M.V., Rokas, A., 2020. Quartet-based computations of internode certainty provide robust measures of phylogenetic incongruence. Syst Biol. 69 (2), 308–324.
Zimmer, E.A., Roalson, E.H., Skog, L.E., Boggan, J.K., Idnurm, A., 2002. Phylogenetic relationships in the Gesnerioideae (Gesneriaceae) based on nrDNA ITS and cpDNA trnL-F and trnE-T spacer region sequences. Am. J. Bot. 89 (2), 296–311.