universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Phylogeny and evolution of Orobanchaceae“

verfasst von / submitted by

Xi Li

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2017 / Vienna 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 794 685 437
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Biologie / Biology
Betreut von / Supervisor:	Assoz.-Prof. Mag. Dr. Gerald M. Schneeweiss

Acknowledgements

I would like to express my gratitude to:

my supervisor - Gerald M. Schneeweiss for responding to my questions in great patience, for providing insightful ideas and discussions, especially for checking my thesis word by word, from which I benefited a lot from writing. I appreciate his time and energy on my work, and his rigorous scholarship was contagious for me. His humorous and accommodating characteristics made my Ph.D. period relaxed and enjoyable. He is a marvellous mentor for me.

my parents – for taking care of themselves well during the time I was not around them, and encouraging me all the time, which enable me to concentrate on my work.

my colleagues – Dr. Hanna Weiss-Schneeweiss for generous sharing chemicals, and her faithful support. Da Pan for fruitful discussion and for providing me the bioinformatics support. Baohai Hao for cooperating in part of my work, and helping me resolve issues of bioinformatics. He is like doraemon, for giving me so many scripts and insightful suggestions out of his magical pocket. Clemens, Hidetoshi, Koji and all the other colleagues for collecting in the field, Tae-Soo, Eva, Elfriede and Michael for their contribution in lab work.

CSC – for supporting from the program of China Scholarships Council (No.201206100007), which covered all the necessary cost in the past three years.

my friends – Xiangjie Ma, Lei Wang, Dan Ma, Lei Gui, and Wangshu Wang for their accompany and encouragement during the past years.

The research of my work was conducted at the Department of Botany and Biodiversity Research of the University of Vienna.

Table of contents

Abstract

Orobanchaceae is a model group for studying the evolution of parasitic flowering plants, because it contains the whole range of nutritional modes from autotrophic non-parasites via photosynthetic parasites to non-photosynthetic parasites. The globally distributed family contains ca. 2060 sepcies in 90 genera, some of which contain pest species causing major yield losses and sinister damages on different crops. Investigating the path to heterotrophic dependence, shifts in host range, and interactions between host and parasite requires a well-supported phylogenetic framework.

Molecular phylogenetic studies have greatly improved our understanding of phylogenetic relationships in Orobanchaceae, but 1/3 of genera have remained unstudied. One of those is the East Asian genus *Platypholis*. Based on overall morphological similarity, *Platypholis* has been merged with *Orobanche*, a hypothesis tested with molecular data in the first chapter of this thesis. Employing maximum likelihood and Bayesian analyses on a family-wide data set (two plastid markers, *matK* and *rps2*, and three nuclear markers, ITS, *phyA* and *phyB*) as well as on an ITS data set focusing on *Orobanche* s. str., it is shown that *P. boninsimae* is phylogenetically nested within *Orobanche* s. str. This is supported by the newly obtained chromosome number of 2*n* = 38 and is congruent with previous morphological evidence.

Although in phylogenetic analysis of ca. 2/3 of the genera in Orobanchaceae six major clades have been identified, the relationships among some major clades are still ill-defined. Consequently, more molecular markers are needed to alleviate these issues, which is the focus of the second chapter of this thesis. Here, five nuclear markers (three low-copy nuclear (LCN) genes and two pentatricopeptide repeat (PPR) genes) are sequenced and analysed using maximum likelihood and Bayesian analysis on a family-wide data set. Compared to previous studies, we additionally included *Macrosyringion*, *Nothobartsia*, *Odontitella*, *Phtheirospermum* (except *Phtheirospermum japonicum*), *Pterygiella*, *Remannia* and *Triaenophora*. The PPR genes used in our study performed well, supporting their utility for studying phylogenies of non-model and parasitic plant groups at intergeneric and interspecific levels, while the used LCN markers are less suitable. Our results strongly suggest that *Pterygiella* and *Phtheirospermum* form a separate clade distinct from the previously recognized major clades. Despite the increased amount of data, we detected several major discordances to previous studies (e.g., concerning the relationships of the *Euphrasia*-*Rhinanthus* clade, the *Castilleja*-*Pedicularis* clade and the *Striga*-*Alectra* clade).

Evidently, two hand-full of loci may still not be sufficient to establish phylogenetic relationships, calling for alternative approaches. The combination of target enrichment with next-generation sequencing (NGS) strategies has the potential to yield a large number of LCN loci and is becoming increasingly popular in phylogenomics. In the third chapter, we establish a highly efficient bioinformatic pipeline to discover putative orthologous single-copy genes (SCGs) for the plant family Orobanchaceae. We used transcriptome data of differing quality available for four Orobanchaceae species and, as reference, SCG data from monkeyflower (1,915 genes) and tomato (391 genes). Depending on whether gaps were permitted in initial BLAST searches of the four Orobanchaceae species against the reference, our pipeline identified 1,307 and 981 SCGs respectively, of length of at least 780 bp. Automated bait sequence construction (using 2× tiling) resulted in 38,156 and 21,856 bait sequences, respectively. In comparison to the recently published MarkerMiner 1.0 pipeline ours identified about 1.6 times as many SCGs (of at least 900 bp length).

Zusammenfassung

Orobanchaceae sind eine Modellgruppe für die Untersuchung der Evolution parasitischer Blütenpflanzen. Ein Verständnis der Evolution von Heterotrophie und den Interaktionen zwischen Wirt und Parasit bedarf eines soliden phylogenetischen Rahmens, welcher trotz enormer Fortschritte in den letzten Jahren noch nicht erreicht ist.

Im ersten Kapitel dieser Dissertation werden molekular-phylogenetische Methoden an einem familienweiten Datensatz aus meheren DNA-Regionen sowie neu erhobene karyologische Daten verwendet, um die Stellung der Gattung *Platypholis* festzustellen. Es stellt sich dabei heraus, dass *Platypholis* phylogenetisch zur Gattung *Orobanche* i. e. S. gehört, was auch durch die Chromosomenzahl von $2n = 38$ unterstütz wird.

Im zweiten Kapitel werden fünf nukleäre low-copy-Gene (LCN-Gene), einschließlich zweier Pentatricopeptid-Repeat (PPR) Gene, an einem familienweiten Datensatz verwendet, um bisherige Hypothesen bezüglich verwandtschaftlicher Beziehungen der Hauptgruppen innerhalb der Orobanchaceae zu testen. Wir können die gute Eignung von PPR-Genen auf unterschiedlichen phylogenetischen Ebenen bestätigen, während andere LCN-Gene unzureichend sind. Trotz erhöhter Datenmenge in Hinblick auf analysierte Marker bleiben größere Diskordanzen zwischen den Markern und zu Ertgebnissen vorheriger Studien bestehen.

Nachdem die Verwendung einer Handvoll Marker scheinbar nicht ausreicht, um die phylogenetischen Beziehungen mit ausreichender Sicherheit zu ermitteln, wird ein phylogenomischer Ansatz, der target enrichment (d.h., die Anreicherung spezifischer Regionen) mit neuesten Sequenziermethoden (next-generation sequencing, NGS) kombiniert, erforderlich sein. Im dritten Kapitel entwickeln wir eine effiziente und flexible Computer-Pipeline, mit deren Hilfe orthologe single-copy-Gene (SCG) bei den Orobanchaceae gefunden werden können. Wir verwenden Transkriptomdaten von vier Arten aus den Orobanchaceae sowie, als Referenz, von Gaulklerblume (*Mimulus guttatus*) und Tomate (*Solanum lycopersicum*). Abhängig von den Einstellungen identifiziert unsere Pipeline 1307 bzw. 981 SCGs, die in 38156 bzw. 21856 Probensequenzen (kurze DNA-Fragemnet mit einer Länge von 120 Basenpaaren, die zur Anreicherung via DNA-Hybridisierung verwendet werden) resultieren. Damit erreicht unsere Pipeline im Vergleich zu einer kürzlich publizierten (MarkerMiner) eine um etwa das 1,6-fach höhere Ausbeute an SCGs.

Introduction

Parasitism and evolution in angiosperms

Photosynthesis, as a primary physiochemical process to convert light energy into chemical energy and produce organic compounds, is one of the hallmarks of plants. Some plants have, however, developed different lifestyles and abandoned photosynthesis partly or completely, often causing reduced phenotypic traits and adapted physiology (Yoshida et al. 2016). For instance, plants with the world's largest flower, whose plastid genome has been possibly lost and undergone massive mitochondrial gene transfer (*Rafflesia*, Rafflesiaceae) (Molina et al.2014), or plants with weird appearance whose plastid genome has been drastically reduced (*Hydnora*, Hydnoraceae) (Naumann et al. 2016). There are around 4,000 parasitic species in angiosperms, and parasitism as an old trait within angiosperms that has independently evolved about twelve times (Naumann et al. 2013). Parasites that live epiphytically on the stem of their host are called shoot parasities (e.g., *Cuscuta*, Convolvulaceae) (Krause 2008), whereas parasites that attach to the host roots they are called root parasities (e.g., *Orobanche*, Orobanchaceae) (Schneeweiss 2013). Given the retention or not of photosynthetic activity, hemi- and holoparasites can be distinguished. Irrespectively if they are root or shoot parasites, hemi- or holoparasites exploit their hosts by a unique organ, the haustorium. Thus, parasites obtain from their hosts water, nutrients, but also genetic material. Due to the special nutritional modes and loss of photosynthesis, parasites have been regarded to be model to study the mechanisms of functional plastid and mitochondrial genome degradation and evolutionary rate variation (Molina et al., 2014; Naumann et al., 2016; Wicke et al., 2016; Fan et al., 2016) and to study horizontal gene transfer (HGT), especially in the mitogenome (Sanchez-Puerta et al., 2008; Xi et al., 2012, 2013; Cusimano and Wicke 2015;Yang et al., 2016).

Phylogenetic and evolutionary trends in Orobanchaceae

Orobanchaceae is the largest parasitic family in angiosperms and comprises c.2060 species in 90 genera, which are of worldwide distribution mainly in the northern hemisphere (McNeal et al., 2013). Notably, some parasitic Orobanchaceae species cause major yield losses and sinister damages on different crops (Heide-Jørgensen 2008). Orobanchaceae have been regarded to be a model group for studying the evolution of parasitic flowering plants, for it is the unique family with whole range of contrasting nutritional modes. In the family parasitism has evolved only once, the transition from hemi- to holoparasitism has occurred at least three times independently, and the family is considered to be monophyletic

(Schneeweiss 2013). Investigating the path of heterotrophic dependence and interactions between host and pest, requires a well-supported phylogenetic framework (Wolfe et al., 2005; Bennett and Mathews 2006). The phylogenetic relationships within Orobanchaceae have been studied by using different plastid and nuclear markers, and certain consensus has been reached concerning delimitation of several major clades. Despite enormous progress with respect to elucidating phylogenetic relationships within Orobanchaceae, there are still conflicts in some clades among different markers (McNeal et al. 2013). A considerable number of genera have not been studied yet using molecular phylogenetic approaches, for instance lacking for the East Asian *Platypholis boninsimae* endemic to Bonin-Islands (SE Japan), rendering our knowledge on phylogenetic relationships in Orobanchaceae incomplete (Schneeweiss 2013). On the other hand, considering the uni-parental inheritance, genome reduction, gene pseudogenization, HGT, and insufficient evolutionary rate, plastid and mitochondrial genes may be of limited suitability. Low-copy nuclear genes (LCN) and PPR genes have been shown to have more advantages in phylogeny studies (Sang 2002; Li et al., 2008; Yuan et al., 2010; Babineau et al., 2013; Crowl et al., 2014).

Research Goals

The primary goal of my thesis aimed at clarifying the phylogenetic relationships in Orobanchaceae by expanding both sampling and molecular loci. The first part of the doctoral thesis aims to test the phylogenetic position of *Platypholis* by employing maximum likelihood and Bayesian analyses on a family-wide data set (including two plastid markers and three nuclear markers) as well as on an ITS data set focusing on *Orobanche* s. str. Additionally, chromosome numbers and genome size data of *Platypholis* will be obtained as well. Over all, chapter 1 will addresses the following questions:

(1) Can the previous morphology-based hypothesis with respect to the phylogenetic position of *Platypholis* within *Orobanche* (Tuyama 1937, 1946) be corroborated?

(2) What are the closest relatives of *Platypholis*?

(3) Are chromosome number and genome size data congruent with the molecular data?

For a better understanding of the phylogenetic relationships at the family level in Orobanchaceae, in chapter 2 we aim at utilizing PPR genes, previously demonstrated to be suitable markers by Yuan et al (2010), and newly developed more LCN loci and expanded sampling. We also compare our results to previous studies in order to evaluate efficiency of the loci we use in resolving intergeneric phylogenetic problems. Furthermore, we aim to obtain a broader family-wide data set with extended sampling (including *Macrosyringion*, *Nothobartsia, Odontitella, Phtheirospermum* (except *Phtheirospermum japonicum*), *Pterygiella, Rehmannia* and *Triaenophora*), to test available phylogenetic hypothesis in Orobanchaceae. Specific questions of chapter 2 include:

(1) Are the phylogenetic relationships of major clades identified by our markers consistent with those found previously by McNeal et al. (2013)?

(2) Are currently recognized clades/genera monophyletic?

Considering that two hands of single loci may still be insufficient to resolve the intricate relationships at different levels, alternative approaches may be needed. Phylogenetics has benefitted from single-copy nuclear genes (SCG) obtainable from next-generation sequencing (NGS) data in the last decade. We aim at establishing a new pipeline to identify putative orthologous SCGs to be used in a target enrichment approach in Orobanchaceae. We also would like to compare our pipeline to MarkerMiner 1.0 and other related studies for phylogenomics, retaining the merits from their pipelines and improving on some aspects to make the pipeline more flexible and suitable also for other non-model plants, especially parasitic plants, where only transcriptomes or partial genome data are available.

References

Babineau M, Gagnon E et al (2013) Phylogenetic utility of 19 low copy nuclear genes in closely related genera and species of caesalpinioid legumes. South African Journal of Botany 89: 94-105.

Bennett,J.R.,S.Mathews (2006) Phylogeny of the parasitic plant family Orobanchaceae inferred from phytochrome A.American Journalof Botany 93 : 1039 – 1051

Crowl AA, Mavrodiev E, Mansion G, Haberle R, Pistarino A, Kamari G, et al. (2014) Phylogeny of Campanuloideae (Campanulaceae) with Emphasis on the Utility of Nuclear Pentatricopeptide Repeat (PPR) Genes. Louis EJ, editor. PLoS ONE 9(4): e94199

Cusimano N, Wicke S.(2015) Massive intracellular gene transfer during plastid genome reduction in nongreen Orobanchaceae[J]. New Phytologist 210:680–693.

Fan W, Zhu A, Kozaczek M, et al. (2016) Limited mitogenomic degradation in response to a parasitic lifestyle in Orobanchaceae [J]. Scientific reports 6.

Li M, Wunder J, Bissoli G, Scarponi E, Gazzani S, Barbaro E, Saedler H, Varotto C (2008) Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species. Cladistics 24,727–745.

Heide-Jørgensen HS (2008) Parasitic flowering plants. Brill, Leiden

Krause, K. (2008). From chloroplasts to “cryptic” plastids: evolution of plastid genomes in parasitic plants. *Current Genetics, 54*(3), 111-121.

Molina, J., Hazzouri, K. M., Nickrent, D., et.al. (2014). Possible loss of the chloroplast genome in the parasitic flowering plant *Rafflesia lagascae* (Rafflesiaceae). Molecular biology and evolution, 31(4), 793-803.

McNeal J, Bennett JR, Wolfe AD, Mathews S (2013) Phylogeny and origins of in Orobanchaceae. Am J Bot 100(5): 971-983.

Naumann J, Salomo K, Der J P, et al.(2013) Single-copy nuclear genes place haustorial Hydnoraceae within Piperales and reveal a Cretaceous origin of multiple parasitic angiosperm lineages[J]. PLoS One 8(11): e79204.

Naumann J, Der J P, Wafula E K, et al.(2016) Detecting and characterizing the highly divergent plastid genome of the nonphotosynthetic parasitic plant *Hydnora visseri* (Hydnoraceae)[J]. Genome biology and evolution 8(2): 345-363.

Sang T (2002) Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit Rev Biochem Mol Biol 37:121–147

Sanchez-Puerta MV, Cho Y, Mower JP, Alverson AJ, Palmer JD. (2008) Frequent, phylogenetically local horizontal transfer of the *cox1* group I intron in flowering plant mitochondria. Mol Biol Evol. 25(8):1762–1777.

Schneeweiss GM (2013) Phylogenetic relationships and evolutionary trends in Orobanchaceae. In: Joel DM, Gressel J, Musselman LJ (eds.) Parasitic Orobanchaceae. Springer, Wien & al.

Tuyama T (1937) On *Platypholis boninsimae* Maximowicz and its systematic position. Bot Mag 51:279–285

Tuyama T (1946) Ogasawara-to Tokusan Shimautsubo nitsuite [*Orobanche boninsimae* endemic to the Ogasawara Islands]. Shigen Kagaku Kenkyusho Iho 10:17–18 [in Japanese]

Wolfe AD, Randle CP, Liu L, Steiner KE (2005) Phylogeny and biogeography of Orobanchaceae. Folia Geobot 40:115–134

Wicke S, Müller KF, dePamphilis CW, Quandt D, Bellot S, Schneeweiss GM. (2016) Mechanistic model of evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants. Proc Natl Acad Sci USA 113:9045–50.

Xi Z, Bradley R K, Wurdack K J, et al. (2012) Horizontal transfer of expressed genes in a parasitic flowering plant [J]. BMC genomics 13(1): 227.

Xi Z, Wang Y, Bradley R K, et al. (2013) Massive mitochondrial gene transfer in a parasitic flowering plant clade [J]. PLoS Genet 9(2)

Yang Z, Zhang Y, Wafula E K, et al.(2016) Horizontal gene transfer is more frequent with increased heterotrophy and contributes to parasite adaptation[J]. Proceedings of the National Academy of Sciences 113(45): E7010-E7019.

Yoshida, S., Cui, S., Ichihashi, Y., & Shirasu, K. (2016). The haustorium, a specialized invasive organ in parasitic plants. Annual review of plant biology,67, 643-667.

Yuan Y.W., Liu C., Marx H.E., Olmstead R.G (2010) An empirical demonstration of using pentatricopeptide repeat (PPR) genes as plant phylogenetic tools: Phylogeny of Verbenaceae and the *Verbena* complex. Molecular Phylogenetics and Evolution 54: 23 – 35.

Chapter 1

Molecular and karyological data confirm that the enigmatic genus *Platypholis* from Bonin-Islands (SE Japan) is phylogenetically nested within *Orobanche* (Orobanchaceae)

Published in Journal of Plant Research

J Plant Res (2017) 130:273–280
DOI 10.1007/s10265-016-0888-y

REGULAR PAPER

Molecular and karyological data confirm that the enigmatic genus *Platypholis* from Bonin-Islands (SE Japan) is phylogenetically nested within *Orobanche* (Orobanchaceae)

Xi Li[1] · Tae-Soo Jang[1] · Eva M. Temsch[1] · Hidetoshi Kato[2] · Koji Takayama[3] · Gerald M. Schneeweiss[1]

Received: 9 February 2016 / Accepted: 26 October 2016 / Published online: 21 December 2016

Abstract Molecular phylogenetic studies have greatly improved our understanding of phylogenetic relationships of non-photosynthetic parasitic broomrapes (*Orobanche* and related genera, Orobanchaceae), but a few genera have remained unstudied. One of those is *Platypholis*, whose sole species, *Platypholis boninsimae*, is restricted to the Bonin-Islands (Ogasawara Islands) about 1000 km southeast of Japan. Based on overall morphological similarity, *Platypholis* has been merged with *Orobanche*, but this hypothesis has never been tested with molecular data. Employing maximum likelihood and Bayesian analyses on a family-wide data set (two plastid markers, *matK* and *rps2*, and three nuclear markers, ITS, *phyA* and *phyB*) as well as on an ITS data set focusing on *Orobanche* s. str., it is shown that *P. boninsimae* Maxim. is phylogenetically closely linked to or even nested within *Orobanche* s. str. This position is supported both by morphological evidence and by the newly obtained chromosome number of $2n=38$, which is characteristic for the genus *Orobanche* s. str.

Keyword Bonin Islands · Chromosome number · Molecular phylogeny · *Orobanche* · Parasitic plant · *Platypholis*

Introduction

Orobanchaceae have become a model group for studying the evolution of parasitic flowering plants (Westwood et al. 2010), because the family includes the full range of nutritional modes (from nonparasitic via photosynthetic parasitic to non-photosynthetic parasitic) as well as a number of pest species parasitic on economically important crop plants (Heide-Jørgensen 2008). For a better understanding of the evolution of parasitism and associated changes, a sound phylogenetic framework is needed. Despite enormous progress with respect to elucidating phylogenetic relationships within Orobanchaceae (McNeal et al. 2013), a considerable number of genera have not been studied yet using molecular phylogenetic approaches (Schneeweiss 2013), rendering our knowledge on phylogenetic relationships in Orobanchaceae incomplete.

The highest diversity of non-photosynthetic parasitic (i.e., holoparasitic) species within Orobanchaceae is found in the exclusively holoparasitic *Orobanche* clade. While relationships and circumscription of its constituent genera are largely established (Schneeweiss 2013), molecular data are still lacking for the two East Asian genera *Phacellanthus* Siebold and Zucc. and *Platypholis* Maxim., the latter the focus of the present study. *Platypholis* contains a single species, *P. boninsimae* Maxim. (Fig. 1). It is endemic to the Bonin-Islands (Ogasawara I.) about 1000 km southeast of Japan, where it grows in shady, moist forests parasitizing mainly *Callicarpa subpubescens* Hook. and Arn. (Tuyama 1937). *Platypholis* was first described by Maximowicz

Electronic supplementary material The online version of this article (doi:10.1007/s10265-016-0888-y) contains supplementary material, which is available to authorized users.

✉ Gerald M. Schneeweiss
gerald.schneeweiss@univie.ac.at

[1] Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Vienna, Austria

[2] Makino Herbarium, Tokyo Metropolitan University, 1-1 Minami-Ohsawa, Hachioji-shi, Tokyo 192-0397, Japan

[3] Museum of Natural and Environmental History, Shizuoka, 5762 Oya, Suruga-ku, Shizuoka-shi, Shizuoka 422-8017, Japan

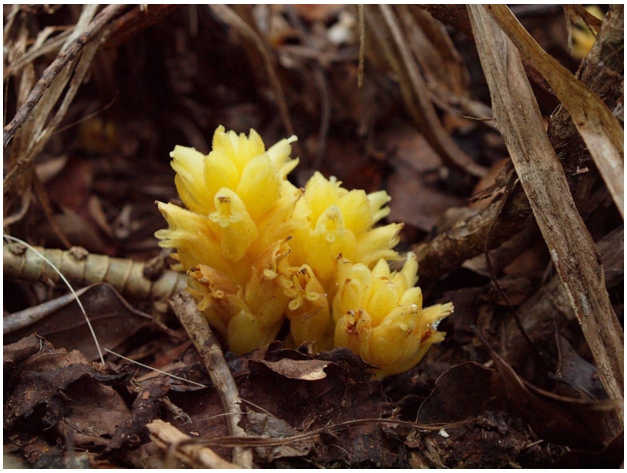


Fig. 1 Habit of *Orobanche boninsimae* (syn. *Platypholis b.*) on Mt. Chibusayama, Hahajima Island (photo by H. Kato)

(1886). He contrasted *Platypholis* with *Conopholis* Wallr., *Boschniakia* C.A.Mey. ex Bong., and *Lathraea* L. (the last not belonging to the *Orobanche* clade: McNeal et al. 2013; Schneeweiss 2013) that differ from *Platypholis* by non-exserted stamens as well as calyx and/or ovary structure. In conflict with Maximowicz's (1886) description, Beck-Manngetta (1890, 1895, 1930) considered *Platypholis* to have three carpels with six placentas (instead of two carpels with four placentas) and consequently put it, together with *Xylanche* Beck (now merged with *Boschniakia* s. str.: Schneeweiss 2013) and *Phacellanthus*, into his Orobanchaceae tricarpellatae. Tuyama (1937) confirmed the observations of Maximowicz (1886) concerning the ovary structure of *Platypholis*. Furthermore, he considered *Platypholis* to be morphologically sufficiently similar to *Orobanche* s. l. to actually merge both genera and treat *P. boninsimae* as *Orobanche boninsimae* (Maxim.) Tuyama (Tuyama 1946). None of these hypotheses has, however, been tested with molecular data yet.

Here we want to clarify the phylogenetic position of *Platypholis* by testing previous hypotheses with respect to the phylogenetic position of *Platypholis* as distinct from *Orobanche* L. (Beck-Mannagetta 1890, 1895, 1930; Maximowicz 1886; Zhang 1988) versus within *Orobanche* (Tuyama 1937, 1946). To this end, we conducted phylogenetic analyses on a family-wide data set (comprising two plastid markers, *matK* and *rps2*, and three nuclear markers, ITS, *phyA* and *phyB*) as well as on an ITS data set focusing on *Orobanche* s. str. (see Schneeweiss 2013, for details on this narrower circumscription of *Orobanche*). As chromosome numbers and genome size data have been found to be phylogenetically informative in the *Orobanche* clade in general and in *Orobanche* s. l. in particular (Schneeweiss et al. 2004b; Weiss-Schneeweiss et al. 2006), these data were obtained as well. Specifically, if *Platypholis* indeed belongs to *Orobanche* (Tuyama 1937, 1946) we expect *Platypholis* to have a chromosome base number of $x = 19$.

Materials and methods

Plant material

Material of *Platypholis* was collected in 2014 in Higashidaira, Chichijima Island, Ogasawara (Bonin) Islands, Japan; the voucher is deposited at WU. For karyological and cytological investigation, young flower buds were fixed in the field in 3:1 ethanol:glacial acetic acid for at least 24 h at room temperature and stored at −20 °C until further use.

DNA extraction, PCR and sequencing

Total DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. Two plastid loci (*matK, rps2*) as well as three nuclear loci (ITS, *phyA* and *phyB*) that have been successfully used in previous phylogenetic studies of Orobanchaceae (McNeal et al. 2013) were amplified using primers listed in Table 1; new or modified primers were designed by eye from available alignments. Amplification of the plastid markers and of ITS was done in a volume of 15.7 µL containing 7 µL KAPA2G Fast2x ReadyMix (Peqlab, Vienna, Austria), 0.5 µL each of 10 µM primer, 1 µL of DNA extract of unknown concentration, and 7 µL sterile water. Amplification of the two phytochrome regions was done in a volume of 15.5 µL containing 0.375 U of Platinum High Fidelity Taq (Invitrogen, Carlsbad, California), 1.5 µL of 10× PCR buffer, 0.8 µL of 50 mM $MgSO_4$, 0.15 µL of 10 mM dNTPs, 0.5 µL each of 10 µM primer, 0.7 µL of DNA extract of unknown concentration and 10.85 µL sterile water. PCR conditions for ITS amplification were: denaturation for 4 min at 94 °C; 35 cycles each with 30 s at 94 °C, 30 s at 48 °C, 1 min at 72 °C; and final elongation for 10 min at 72 °C. For the remaining four loci (*rps2, matK, phyA* and *phyB*) a touchdown PCR protocol was used, thus obviating potential problems due to degenerate primers. The PCR conditions were: 2 min at 94 °C; 9 cycles each with 30 s at 94 °C, 15 s at 67 °C (decreasing the annealing temperature by 1 °C at each subsequent cycle, so that in the 9th cycle the annealing temperature was 59 °C), 90 s at 70 °C; 21 cycles each with 30 s at 94 °C, 30 s at 57 °C, 90 s at 70 °C; 12 cycles with 30 s at 94 °C, 45 s at 62 °C, 90 s at 70 °C; a final elongation for 7 min at 70 °C. PCR products were purified using Exonuclease I and FastAP thermosensitive alkaline phosphatase (Fisher Scientific, St. Leon-Rot, Germany) following the manufacturer's instructions. Cycle sequencing reactions were performed using 5 µL of purified

Table 1 Amplification primers

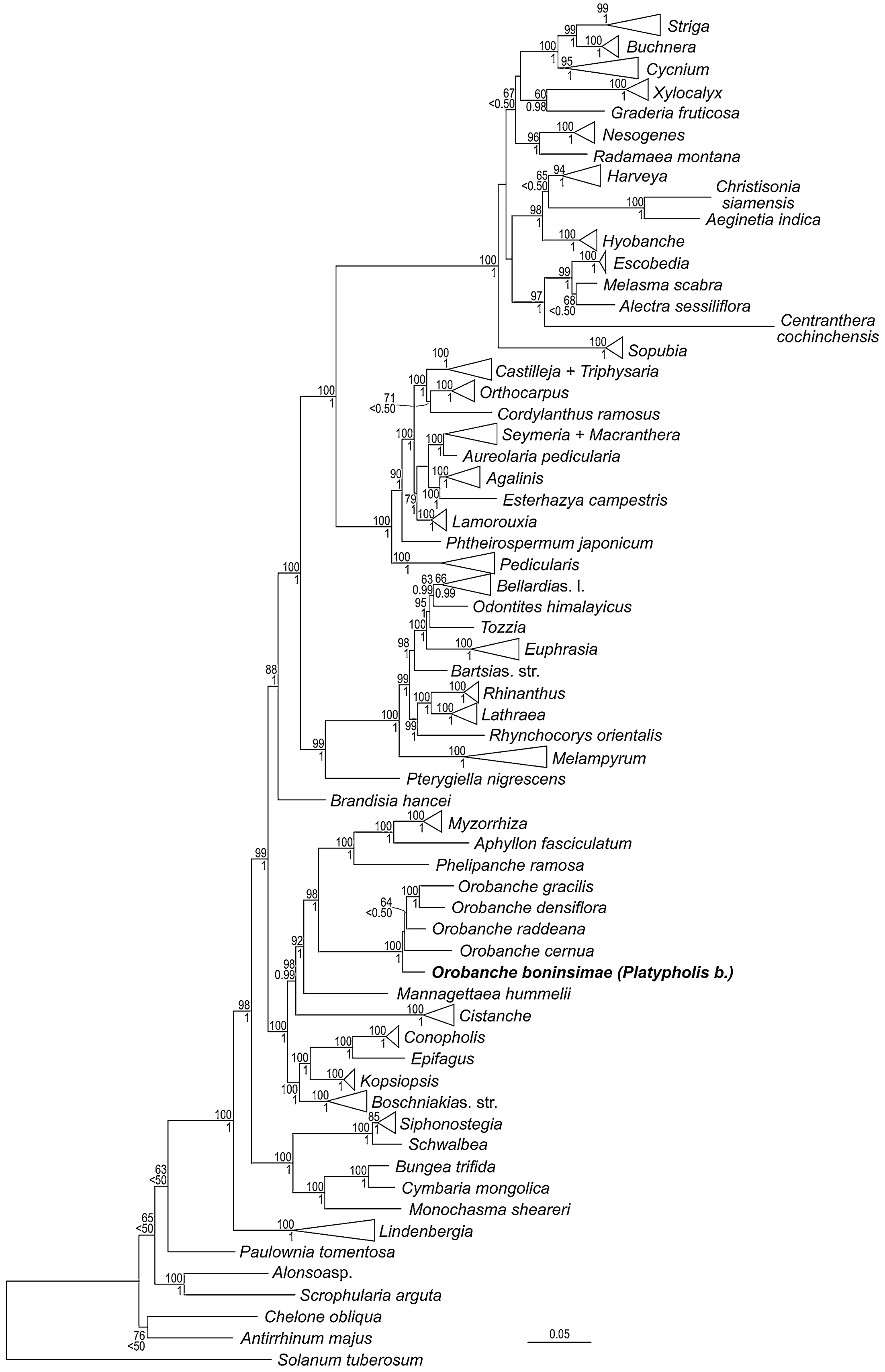
Primer		References
Name	Sequence (5′–3′)	
matK		
trnK 3914F di	GGGGTTGCTAACTCAACGG	Johnson and Soltis (1995)
matK550Fca	TGGAAATCTTGGTTCAAACTCTTCG	This study
matK-50Fdi	GTTTTGACTGTATCGCACTATGTATC	Demaio et al. (2011)
matK 950r	CCACARCGAAAAATRMCATTGCC	Young et al. (1999)
matK 1349r	CTTTTGTGTTTCCGAGCYAAAGTTC	Young et al. (1999)
trnK-R2*	CTCGAACCCGGAACTAGTCGG	Castello et al. (2016)
rps2		
rps2-47F	CTCGTTTTTTATCTGAAGCCTG	dePamphilis et al. (1997)
rps2-58F	AAATGGAATCCTAAAATGGCA	This study
rps2-661R	ACCCTCACAAATAGCGAATACCAA	dePamphilis et al. (1997)
ITS		
ITS AB101	ACGAATTCATGGTCCCGTGAAGTGTTCG	Schneeweiss et al. (2004a)
ITS AB102	TAGAATTCCCCGGTTCGCTCGCCGTTAC	Schneeweiss et al. (2004a)
phyA		
PHYA230f	GACTTTGARCCNGTBAAGCCTTAYG	Mathews and Donoghue (1999)
PHYA_Newa678r	GTCTCRATCARACGAACCATCTC	This study
phyB		
PHYB7f	CACAGGATAGAYGTRGGRGT	This study
NewPHYB_b678r.oro	GTCTCTATCAACCTAAYCATCTC	This study (modified from McNeal et al. 2013)

template, 1 µL of primer (3.2 µM) and 1 µL BigDye Terminator (Applied Biosystems, Foster City, California), cleaned with Sephadex G-50 Fine (GE Healthcare Bio-Sciences, Uppsala, Sweden) and sequenced on an ABI 3730 DNA Analyzer capillary sequencer (Applied Biosystems).

Phylogenetic analyses

Sequences were assembled and edited using SeqMan II 5.05 (DNAStar Inc., Madison, USA). The newly obtained data of *Platypholis* were added and aligned by eye to the existing single and combined marker alignments of McNeal et al. (2013), available from TreeBase (http://treebase.org) under study number 13942, using BioEdit 7.2.1 (Hall 1999). Likewise, ITS sequences of *Platypholis* were added to the alignment of Frajman et al. (2013) that focuses on *Orobanche* s. str. and consequently has a much denser sampling within that genus. Sequence alignments are available from ResearchGate under doi:10.13140/RG.2.1.4124.1203.

The best-fit substitution models were identified using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as implemented in jModelTest 2.1.6 (Darriba et al. 2012). We tested 44 substitution models (11 substitution schemes, allowing unequal frequencies and/or rate heterogeneity across sites modeled by a gamma distribution, but no proportion of invariable sites due to identifiability issues: Yang 2014) on maximum-likelihood (ML) optimized topologies obtained following SPR (Subtree Pruning and Regrafting) branch swapping. For each dataset, the General Time Reversible (GTR) model (Tavaré 1986) including rate heterogeneity across sites described by a gamma distribution was selected. Maximum likelihood analyses were conducted using RAxML 8.1 (Stamatakis et al. 2014) employing the fast bootstrap approach (Stamatakis et al. 2008) with 1000 bootstrap replicates. Bayesian inference was done using MrBayes 3.2.3 (Ronquist et al. 2012). Values for all parameters, such as the shape of the gamma distribution (approximated using six discrete rate categories) or the substitution rates, were estimated during the analysis. For partitioned analyses (combined data set only), partitions were allowed to evolve under different rates (ratepr = variable). Four Monte Carlo Markov (MCMC) chains were run simultaneously starting from different random starting trees for 20 million generations, with trees sampled every 5000th generation. After combining 3600 trees from each run (i.e., after discarding 10% of samples as burn-in, when the MCMC chain had reached stationarity as confirmed by visual inspection of traces and standard deviations of split variances being below 0.01), posterior probabilities were estimated from these 14,400 posterior trees and were plotted on a majority rule consensus tree.


99
1
Striga
99
1
100
1
Buchnera
100
1
95
1
Cycnium
67
<0.50
60
0.98
100
1
Xylocalyx
Graderia fruticosa
96
1
100
1
Nesogenes
Radamaea montana
65
<0.50
94
1
Harveya
Christisonia
siamensis
100
1
Aeginetia indica
98
1
100
1
Hyobanche
100
1
Escobedia
99
1
Melasma scabra
68
<0.50
Alectra sessiliflora
97
1
Centranthera
cochinchensis
100
1
100
1
Sopubia
100
1
Castilleja + Triphysaria
100
1
100
1
Orthocarpus
71
<0.50
Cordylanthus ramosus
100
1
100
1
Seymeria + Macranthera
Aureolaria pedicularia
90
1
100
1
Agalinis
79
1
100
1
Esterhazya campestris
100
1
Lamorouxia
100
1
Phtheirospermum japonicum
100
1
Pedicularis
63
0.99
66
0.99
Bellardias. l.
95
1
Odontites himalayicus
100
1
Tozzia
98
1
100
1
Euphrasia
Bartsias. str.
99
1
100
1
100
1
Rhinanthus
100
1
Lathraea
100
1
99
1
Rhynchocorys orientalis
100
1
Melampyrum
99
1
Pterygiella nigrescens
100
1
88
1
Brandisia hancei
100
1
100
1
Myzorrhiza
Aphyllon fasciculatum
100
1
Phelipanche ramosa
99
1
100
1
Orobanche gracilis
Orobanche densiflora
64
<0.50
Orobanche raddeana
98
1
100
1
Orobanche cernua
Orobanche boninsimae (Platypholis b.)
92
1
98
0.99
Mannagettaea hummelii
100
1
Cistanche
98
1
100
1
100
1
100
1
Conopholis
Epifagus
100
1
100
1
Kopsiopsis
100
1
100
1
Boschniakias. str.
85
1
Siphonostegia
100
1
Schwalbea
100
1
100
1
Bungea trifida
Cymbaria mongolica
100
1
Monochasma sheareri
100
1
Lindenbergia
63
<50
Paulownia tomentosa
65
<50
100
1
Alonsoasp.
Scrophularia arguta
Chelone obliqua
76
<50
Antirrhinum majus
0.05
Solanum tuberosum

◂**Fig. 2** Phylogenetic placement of *Orobanche boninsimae* (syn. *Platypholis b.*) within Orobanchaceae inferred using maximum likelihood on a five marker combined data set (*matK*, rps2, ITS, *phyA*, *phyB*). *Numbers* at branches are maximum likelihood bootstrap support values (60 or higher) and posterior probabilities (0.5 or higher). *Solanum tuberosum* L. was chosen as outgroup

Chromosome number and genome size

The chromosome number of *Platypholis* was determined from meiotic divisions in pollen mother cells (PMCs) and from first mitosis in developing microspores using the standard Feulgen staining technique (Schneeweiss et al. 2004b). Fixed material was hydrolyzed in 5N HCl for 30 min at room temperature, washed with tap water and stained with Schiff's reagent (Merck, Darmstadt, Germany) in darkness for 1 h (Jang et al. 2013). Chromosome spreads were prepared by squashing stained anthers in a drop of acetic acid (60%) under the cover-slip, and analyzed using an AxioImager M2 microscope (Carl Zeiss, Vienna, Austria). Preparations with a minimum of 15 good quality chromosome spreads were analyzed. Images were acquired with a CCD camera and files processed using AxioVision 4.8 (Carl Zeiss). The karyotype was made from these images in PhotoPaint X7 (Corel Corp., Ottawa, Ontario).

Fixed flower buds were transferred to ethanol and stored in the deep freezer. For preparation for genome size estimation, plant tissue was rehydrated and hydrolyzed for 60 min in 5N HCl at 20 °C (Greilhuber and Temsch 2001) together with root tips from the internal standard (*Pisum sativum* L. 'Kleine Rheinländerin', 1C = 4.42 pg: Greilhuber and Ebert 1994). After washing with water, the samples were stained with Schiff´s reagent over-night in the refrigerator. The dye was removed by washing six times with SO_2-water over a total period of 45 min. Subsequently, the stained tissue was squashed on slides, frozen, and after removal of cover slips fixed in 96% ethanol, dried and stored until measurement. Measurements of the Integrated Optical Densities (IOD) were conducted on the Cell Image Retrieval and Evaluation System (CIRES, Kontron, Munich), which was equipped with a CCD DXC 390P camera (Sony, Tokyo, Japan) and an Axioscope microscope (Carl Zeiss). From each slide, for both the object and the internal standard, 10 prophase and 10 telophase nuclei were measured. Per slide a 1C-value was calculated using the formula (mean $\mathrm{IOD_{Obj}}$/mean $\mathrm{IOD_{Std}}$)*1C-$\mathrm{value_{Std}}$.

Results

Newly obtained sequences of *matK, rps2*, ITS, *phyA* and *phyB* are available from GenBank under accession numbers KU647699, KU647702, KU647698, KU647700 and KU647701, respectively. Phylogenetic analyses of the family-wide data sets (single marker and concatenated data sets) congruently place *Platypholis* within the *Orobanche* clade as sister to or nested within *Orobanche* s. str. (BS = 100, PP = 1; Online Resource 1; Fig. 2), but low resolution and/or insufficient support (except for ITS: Fig. 3; Fig. S3 in Online Resource 1) prevent the precise phylogenetic position of *Platypholis* being identified (Fig. 2; Online Resource 1). Although phylogenetic relationships among lineages inferred from single markers are not fully congruent (Online Resource 1), well-supported incongruences involving *Platypholis* are lacking and those involving the remaining taxa have been found to be not statistically significant (McNeal et al. 2013). Analyses of a data set focusing on *Orobanche* s. str. place *Platypholis* firmly within *Orobanche* s. str. (BS = 87, PP = 1.00, Fig. 3), where it is inferred as sister species to *O. coerulescens* Stephan (BS = 74, PP = 0.97, Fig. 3).

Platypholis is diploid with a chromosome number of $2n = 2x = 38$ (Fig. 4). All chromosomes are metacentric to submetacentric and their lengths range from 2 to 5 µM (Fig. 4), resulting in a Haploid Karyotype Length (HKL) of 54.83 µM. The nuclear DNA amount (1C) of *Platypholis*, calculated as average from four slide pairs, is 7.28 pg (S.D. 0.1805, C.V. 2.48%).

Discussion

The monotypic genus *Platypholis* has not been included in any molecular phylogenetic study of Orobanchaceae to date and its precise placement within the family remained uncertain (Schneeweiss 2013). Using data from two plastid and three nuclear loci, it is shown that *Platypholis* phylogenetically belongs to the *Orobanche* clade (Fig. 2, Online Resource 1) and is sister to or even nested within *Orobanche* s. str. (Fig. 3). The uncertainty concerning the precise placement of *Platypholis* may be due to issues of paralogy in nuclear markers, especially ITS (Álvarez and Wendel 2003). As neither gel visualization of PCR products nor direct sequencing indicated any presence of paralogues and as there are no strongly supported, but rather contradicting phylogenetic relationships inferred from plastid versus nuclear markers (Online Resource 1), we consider it unlikely that our inferences are misled by paralogues. Alternatively, incongruences between different markers might be due to incomplete lineage sorting, which can be accommodated by using species tree estimation methods. This will, however, require much larger data sets, especially if a possible negative affect of missing data is to be avoided (Xi et al. 2016), which goes beyond the scope of this study.

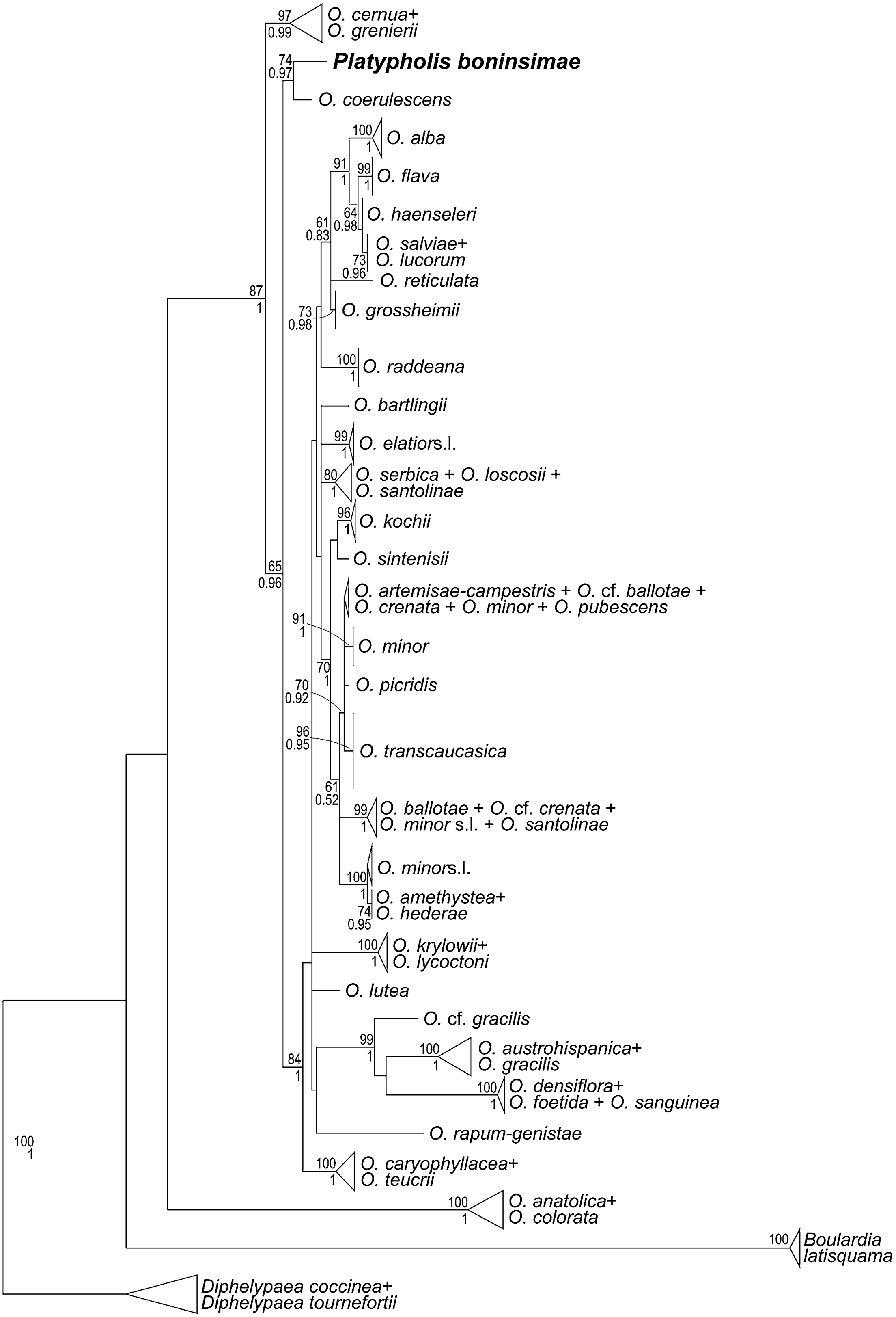


Fig. 3 Phylogenetic placement of *Orobanche boninsimae* (syn. *Platypholis b.*) within *Orobanche* s. str. (i.e., also excluding *Boulardia*: Schneeweiss 2013) inferred using maximum likelihood on an ITS data set. *Numbers* at branches are maximum likelihood bootstrap support values (60 or higher) and posterior probabilities (0.5 or higher). *Diphelypaea* Nicolson was chosen as outgroup (Schneeweiss et al. 2004a)

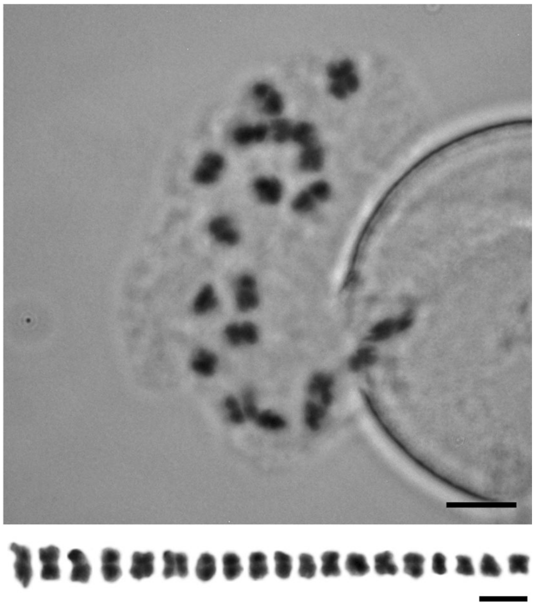


Fig. 4 Chromosomes and karyotype of *Orobanche boninsimae* (syn. *Platypholis b.*): $n = 19$ (metaphase of first mitotic division in microspore). *Scale bar* 5 µm

A close relationship of *Platypholis* and *Orobanche* s. str. is also supported by the shared chromosome number of $2n = 38$ (Fig. 4; Schneeweiss et al. 2004b). Hence, both molecular phylogenetic and karyological data refute (implicit) hypotheses of Maximowicz (1886) and Beck-Mannagetta (1890, 1895, 1930) on a closer relationship to *Lathraea, Conopholis*, and/or *Boschniakia* (no data available for *Phacellanthus*) and instead support Tuyama (1937, 1946), who suggested a close relationship to *Orobanche*. Tuyama (1937) also noted several morphological characters that *Platypholis* shares with all or at least some species of *Orobanche* s. str., including the absence of bracteoles, the flowers being sessile, the calyx being divided into two lateral sepals, the basal insertion of stamens, and the ovary structure (two-carpellate ovaries with four placentae). The last is of particular relevance, because Beck-Mannagetta (1890, 1895, 1930) placed *Platypholis* in his Orobanchaceae tricarpellatae based on the perceived presence of three carpels and six separate placentae (see Fig. 24G in Beck-Mannagetta 1930: 331), while he classified *Orobanche* s. str. (as *O.* sect. *Ospreolon*) within his Orobanchaceae bicarpellatae, due to the presence of two carpels and four separate placentae. Beck's observations are even more puzzling, because Maximowicz (1886), when describing *Platypholis*, had already indicated the presence of four placentae only, which Beck dutifully reported, albeit with reservations (Beck-Mannagetta 1930: 332 "sec. Maximowicz solum 4": "according to Maximowicz only 4"). Taxonomically, the genus *Platypholis* can no longer be upheld and, following Tuyama (1937, 1946) and subsequent Japanese authors, its single species is to be treated as member of *Orobanche* s. str. as *O. boninsimae*.

The phylogenetic placement of *O. boninsimae* within *Orobanche* s. str. is less certain. A closer relationship to *O. coerulescens*, as suggested by ITS data (Fig. 3), is supported by geographic proximity, as *O. coerulescens* (lacking from the Bonin Islands) is the sole *Orobanche* species occurring on the main islands of Japan (Shimane). *Orobanche boninsimae* differs from *O. coerulescens* and other *Orobanche* species by having exserted stamens (unique within the genus; Fig. 1), a stem that is strongly branched at the caudex, larger chromosomes (2–5 µM vs. 1–3 µM: Schneeweiss et al. 2004b) and a correspondingly larger genome (7.28 pg/1C vs. 1.45–5.83 pg/1C: Weiss-Schneeweiss et al. 2006). The considerably larger genome observed in *O. boninsimae* may be connected to life-history (the species is perennial: Tuyama 1937; reports by Abe (2006) that *O. boninsimae* is annual are incorrect) or changes in breeding system, as noted in other plant groups (Albach and Greilhuber 2004; Price et al. 2005).

Acknowledgements Open access funding provided by University of Vienna. We thank Sarah Mathews (Australian National University) for help with amplification protocols for phytochrome genes. Financial support from the program of China Scholarships Council (No. 201206100007) to X. L. is gratefully acknowledged.

References

Abe T (2006) Threatened pollination systems in native flora of the Ogasawara (Bonin) Islands. Ann Bot 98:317–334. doi:10.1093/aob/mcl117

Albach DC, Greilhuber J (2004) Genome size variation and evolution in *Veronica*. Ann Bot 94:897–911. doi:10.1093/aob/mch219

Álvarez I, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. Mol Phylogen Evol 29:417–434. doi:10.1016/S1055-7903(03)00208-2

Beck-Mannagetta G (1890) Monographie der Gattung *Orobanche*. Theodor Fischer, Cassel

Beck-Mannagetta G (1895) Orobanchaceae. In: Engler A, Prantl K (eds) Die natürlichen Pflanzenfamilien IV 3b. Wilhelm Engelmann, Leipzig, pp 123–132. doi:10.5962/bhl.title.4635

Beck-Mannagetta G (1930) Orobanchaceae. In: Engler A (ed) Das Pflanzenreich IV 261. Wilhelm Engelmann, Leipzig, pp 1–348

Castello LV, Barfuss MHJ, Till W, Galetto L, Chiapella JO (2016) Disentangling the *Tillandsia capillaris* complex: phylogenetic relationships and taxon boundaries in Andean populations. Bot J Linn Soc 181(3):391–414. doi:10.1111/boj.12400

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Meth 9:772–772. doi:10.1038/nmeth.2109

Demaio PH, Barfuss MHJ, Kiesling R, Till W, Chiapella JO (2011) Molecular phylogeny of *Gymnocalycium* (Cactaceae): assessment of alternative infrageneric systems, a new subgenus, and trends in the evolution of the genus. Am J Bot 98:1841–1854. doi:10.3732/ajb.1100054

dePamphilis CW, Young ND, Wolfe AD (1997) Evolution of plastid gene *rps2* in a lineage of hemiparasitic and holoparasitic plants: many losses of photosynthesis and complex patterns of rate variation. Proc Natl Acad Sci USA 94:7367–7372

Frajman B, Carlón L, Kosachev P, Sánchez Pedraja O, Schneeweiss GM, Schönswetter P (2013) Phylogenetic position and taxonomy of the enigmatic *Orobanche krylowii* (Orobanchaceae), a predominatly Asian species newly found in Albania (SE Europe). Phytotaxa 137(1):1–14. doi:10.11646/phytotaxa.137.1.1

Greilhuber J, Ebert I (1994) Genome size variation in *Pisum sativum*. Genome 37:646–655

Greilhuber J, Temsch EM (2001) Feulgen densitometry: some observations relevant to best practice in quantitative nuclear DNA content determination. Acta Bot Croat 60:285–298

Hall TA (1999) Bioedit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95–98

Heide-Jørgensen HS (2008) Parasitic flowering plants. Brill, Leiden

Jang TS, Emadzade K, Parker J, Temsch EM, Leitch AR, Speta F, Weiss-Schneeweiss H (2013) Chromosomal diversification and karyotype evolution of diploids in the cytologically diverse genus *Prospero* (Hyacinthaceae). BMC Evol Biol 13(1):136. doi:10.1186/1471-2148-13-136

Johnson LA, Soltis DE (1995) Phylogenetic inference in Saxifragaceae sensu stricto and *Gilia* (Polemoniaceae) using *matK* sequences. Ann Missouri Bot Gard 82:149–175. doi:10.2307/2399875

Mathews S, Donoghue MJ (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science 286:947–950. doi:10.1126/science.286.5441.947

Maximowicz CJ (1886) Diagnoses plantarum novarum asiaticarum. VI.—Insunt stirpes quaedam nuper in Japonia detectae. Bull Acad Imp Sci St 31:12–121

McNeal JR, Bennett JR, Wolfe AD, Mathews S (2013) Phylogeny and origins of holoparasitism in Orobanchaceae. Am J Bot 100:971–983. doi:10.3732/ajb.1200448

Price HJ, Dillon SL, Hodnett G, Rooney WL, Ross L, Johnston JS (2005) Genome evolution in the genus *Sorghum* (Poaceae). Ann Bot 95:219–227. doi:10.1093/aob/mci015

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542. doi:10.1093/sysbio/sys029

Schneeweiss G (2013) Phylogenetic relationships and evolutionary trends in Orobanchaceae. In: Joel DM, Gressel J, Mussleman LJ (eds) Parasitic Orobanchaceae. Springer, Berlin Heidelberg, pp 243–265

Schneeweiss GM, Colwell AE, Park JM, Jang CG, Stuessy TF (2004a) Phylogeny of holoparasitic *Orobanche* (Orobanchaceae) inferred from nuclear ITS sequences. Mol Phylogen Evol 30:465–478. doi:10.1016/S1055-7903(03)00210-0

Schneeweiss GM, Palomeque T, Colwell AE, Weiss-Schneeweiss H (2004b) Chromosome numbers and karyotype evolution of holoparasitic *Orobanche* (Orobanchaceae) and related genera. Am J Bot 91:439–448. doi:10.3732/ajb.91.3.439

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. doi:10.1093/bioinformatics/btu033

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML web servers. Syst Biol 57:758–771. doi:10.1080/10635150802429642

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura RM (ed) Some mathematical questions in biology—DNA sequence analysis (Lectures on Mathematics in the Life Sciences 17). American Mathematical Society, Providence, pp 57–86

Tuyama T (1937) On *Platypholis boninsimae* Maximowicz and its systematic position. Bot Mag 51:279–285

Tuyama T (1946) Ogasawara-to Tokusan Shimautsubo nitsuite [*Orobanche boninsimae* endemic to the Ogasawara Islands]. Shigen Kagaku Kenkyusho Iho 10:17–18 (**In Japanese**)

Weiss-Schneeweiss H, Greilhuber J, Schneeweiss GM (2006) Genome size evolution in holoparasitic Orobanche (Orobanchaceae) and related genera. Am J Bot 93:148–156. doi:10.3732/ajb.93.1.148

Westwood JH, Yoder JI, Timko MP, dePamphilis CW (2010) The evolution of parasitism in plants. Trends Pl Sci 15:227–235. doi:10.1016/j.tplants.2010.01.004

Xi Z, Liu L, Davis CC (2016) The impact of missing data on species tree estimation. Mol Biol Evol 33:838–860. doi:10.1093/molbev/msv266

Yang Z (2014) Molecular evolution. A statistical approach. Oxford Univ Press, Oxford

Young ND, Steiner KE, dePamphilis CW (1999) The evolution of parasitism in Scrophulariaceae/Orobanchaceae: Plastid gene sequences refute an evolutionary transition series. Ann Miss Bot Garden 86:876–893

Zhang ZY (1988) Taxonomy of the Chinese *Orobanche* and its relationships with related genera. Acta Phytotaxon Sin 26:394–403 (**In Chinese**)

Supplementary material

Title: Molecular and karyological data confirm that the enigmatic genus *Platypholis* from Bonin-Islands (SE Japan) is phylogenetically nested within *Orobanche* (Orobanchaceae)

Journal: Journal of Plant Research

Authors: Xi Li, Tae-Soo Jang, Eva M. Temsch, Hidetoshi Kato, Koji Takayma, Gerald M. Schneeweiss

Corresponding author: Gerald M. Schneeweiss, Department of Botany and Biodiversity, University of Vienna, Rennweg 14, A-1030 Vienna, Austria. Fax: +43 1 4277 9541.E–mail: gerald.schneeweiss@univie.ac.at

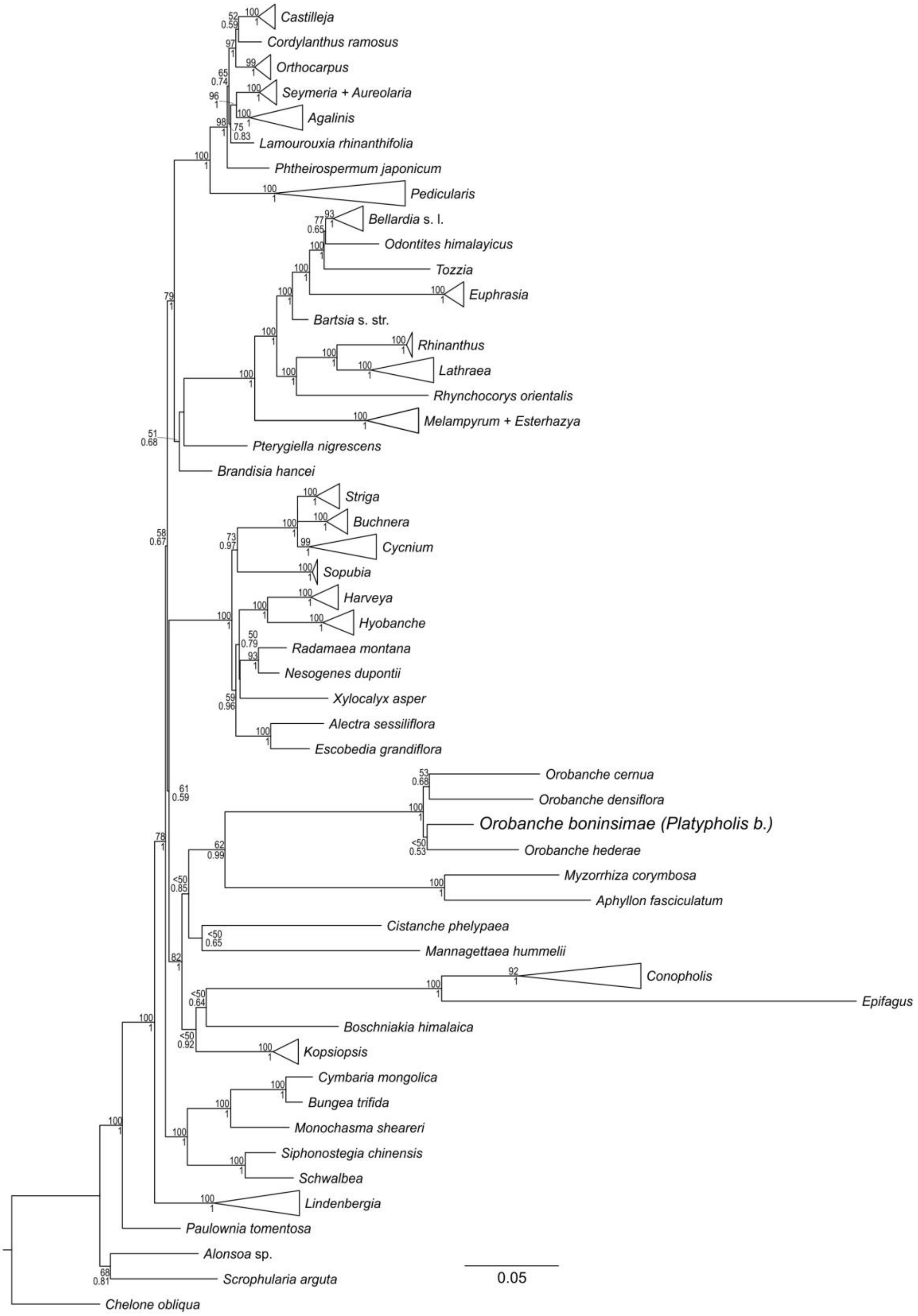


Fig. S1 Phylogenetic placement of *Orobanche boninsimae* (syn. *Platypholis b.*) within Orobanchaceae inferred using maximum likelihood on a *matK* data set. Numbers at branches are maximum likelihood bootstrap support values (50 or higher) and posterior probabilities (0.5 or higher).

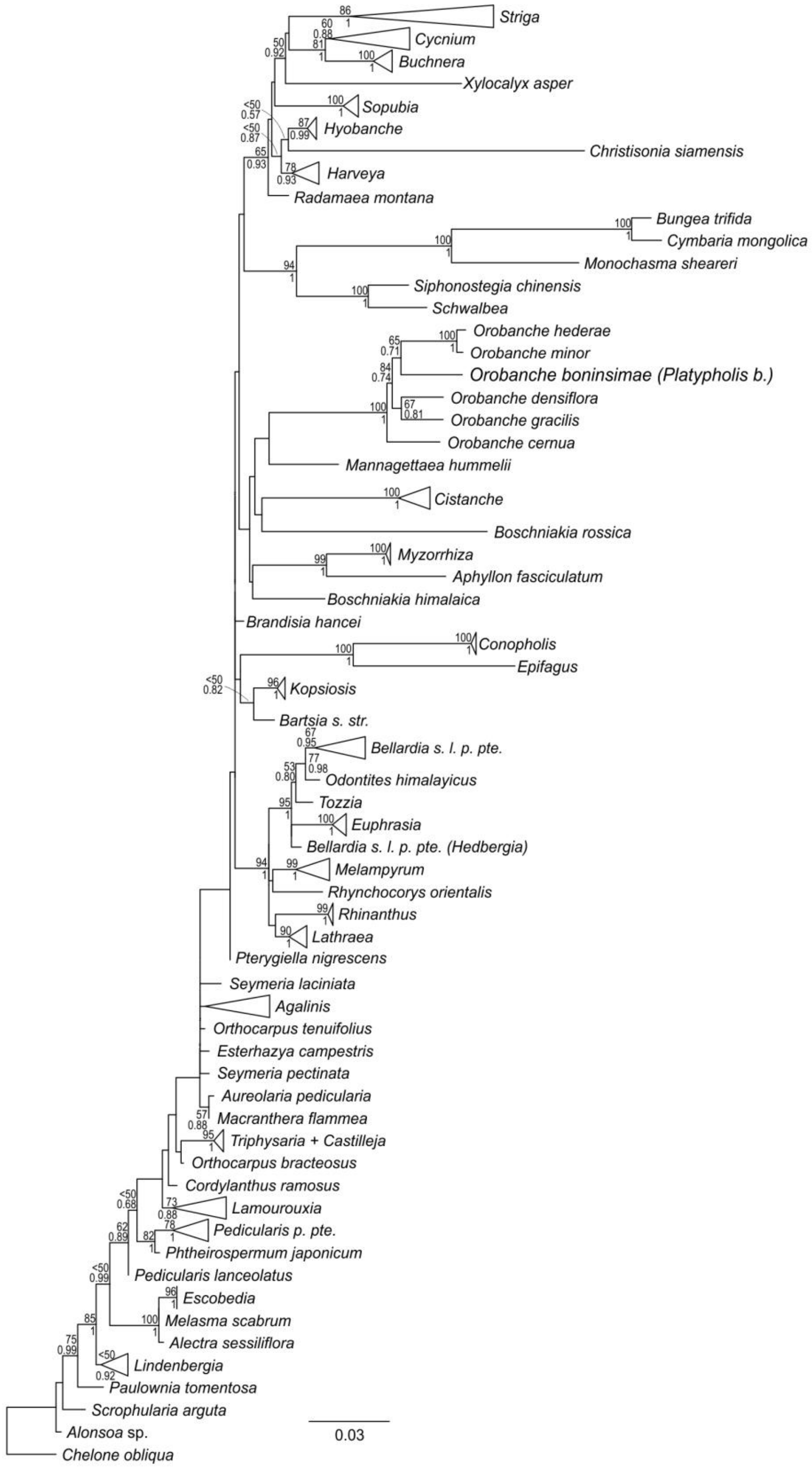


Fig. S2 Phylogenetic placement of *Orobanche boninsimae* (syn. *Platypholis b.*) within Orobanchaceae inferred using maximum likelihood on a *rps2K* data set. Numbers at branches are maximum likelihood bootstrap support values (50 or higher) and posterior probabilities (0.5 or higher).

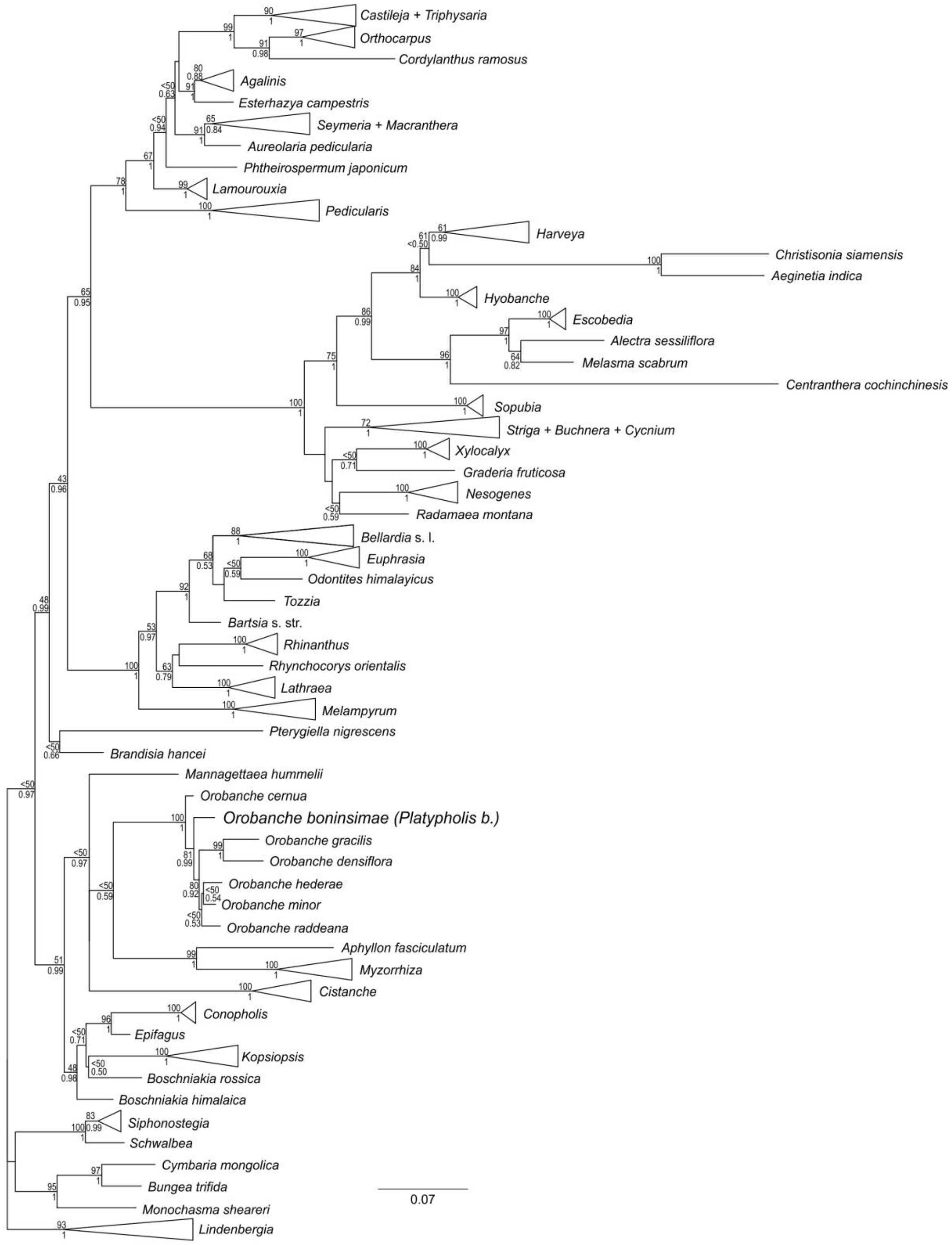


Fig. S3 Phylogenetic placement of *Orobanche boninsimae* (syn. *Platypholis b.*) within Orobanchaceae inferred using maximum likelihood on a ITS data set. Numbers at branches are maximum likelihood bootstrap support values (50 or higher) and posterior probabilities (0.5 or higher).

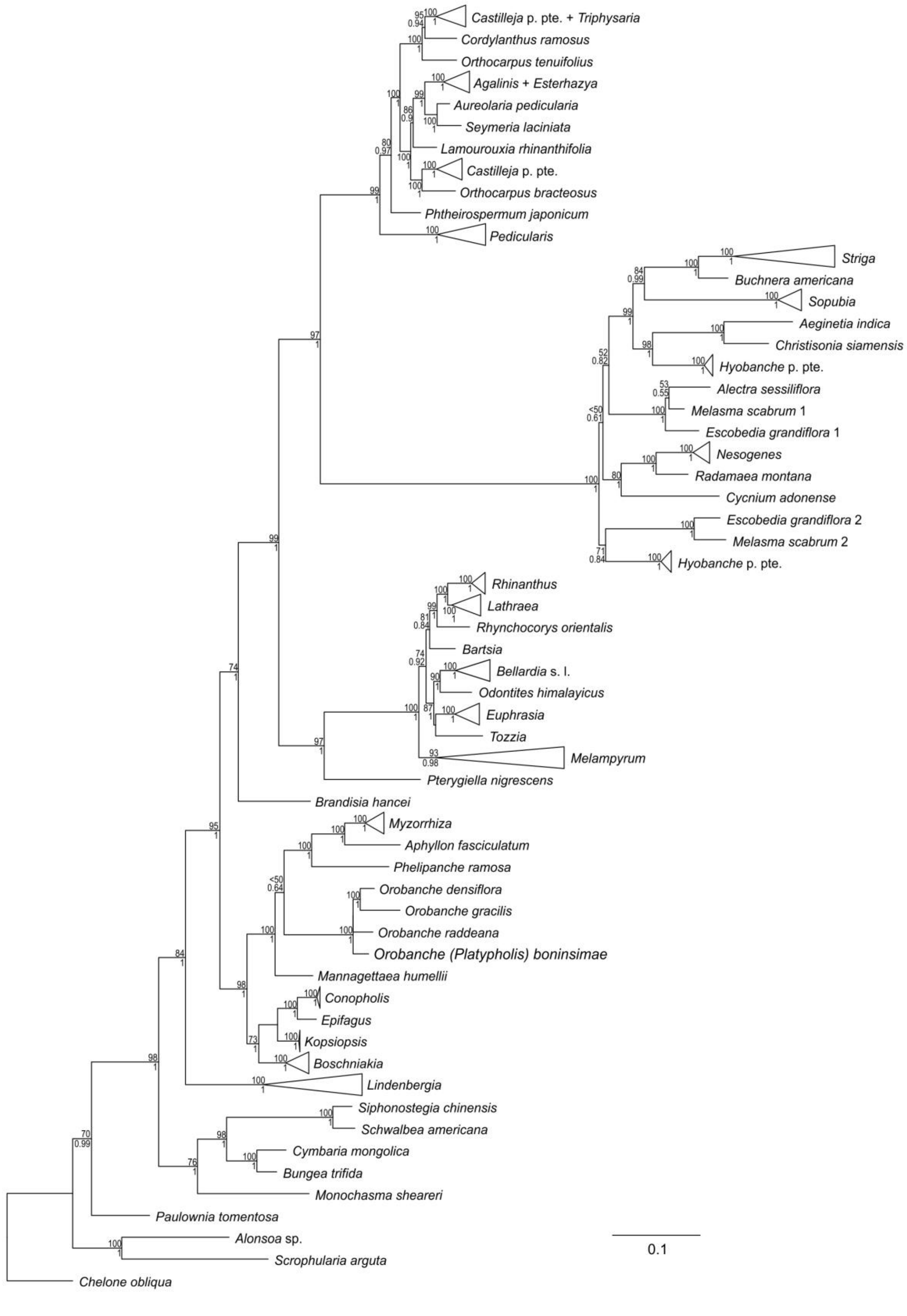


Fig. S4 Phylogenetic placement of *Orobanche boninsimae* (syn. *Platypholis b.*) within Orobanchaceae inferred using maximum likelihood on a *phya* data set. Numbers at branches are maximum likelihood bootstrap support values (50 or higher) and posterior probabilities (0.5 or higher).

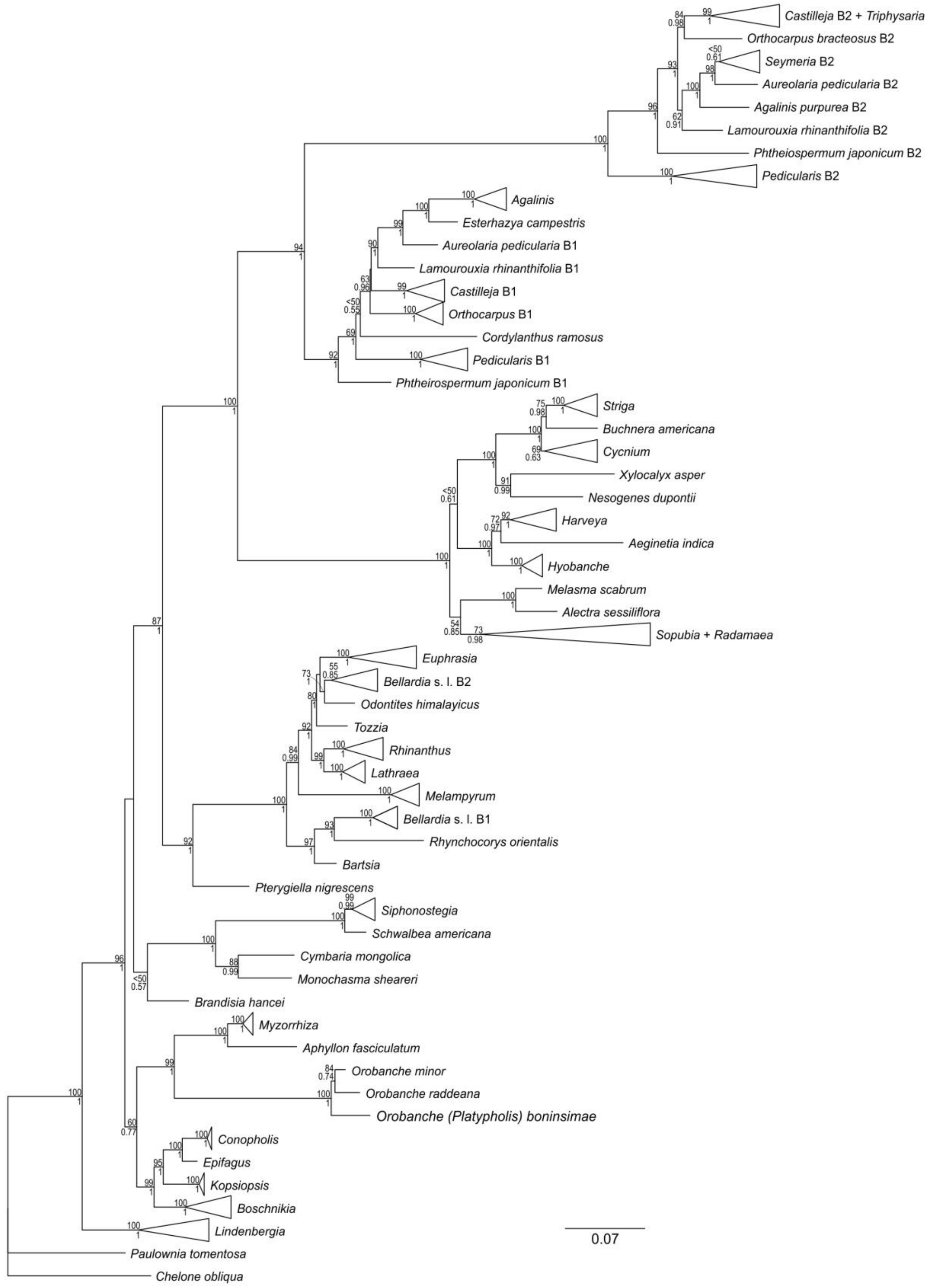


Fig. S5 Phylogenetic placement of *Orobanche boninsimae* (syn. *Platypholis b.*) within Orobanchaceae inferred using maximum likelihood on a *phyb* data set. Numbers at branches are maximum likelihood bootstrap support values (50 or higher) and posterior probabilities (0.5 or higher).

Chapter 2

Phylogenetic utility of pentatricopeptide repeat (PPR) and low copy nuclear (LCN) genes in Orobanchaceae, with whole range of parasitic lifestyle

Phylogenetic utility of pentatricopeptide repeat (PPR) and low copy nuclear (LCN) genes in Orobanchaceae, with whole range of parasitic lifestyle

Xi Li[1], Tao Feng[2], Gerald M. Schneeweiss[1]

[1]Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Vienna, Austria

[2]Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China

Abstract

Considering the unclear relationships within the family in the previous phylogenetic study of Orobanchaceae, we long to develop more powerful orthologous nuclear loci to accurately resolve the backbone relationships. Pentatricopeptide repeat (PPR) and low copy nuclear (LCN) genes have been recently utilized as tools in phylogenomics, profiting from available genomic data. Here, we utilized, for the first time, two PPR genes and three LCN genes employing maximum likelihood and Bayesian analyses at family level of Orobanchaceae. The monophyly of each major clade (except *Pterygiella* and *Phtheirospermum*) was strongly supported here. Although there were minor conflicts among different markers in some of the major clades, these were poorly supported. Furthermore, we detected two major discordances with the previous studies: one was the relationships among *Striga-Alectra* clade, *Euphrasia-Rhinanthus* clade and *Pterygiella* and *Phtheirospermum*, and another was the placement of the holoparasitic *Orobanche* clade. The PPR genes used in our study confirmed the outstanding achievement of PPR gene family in non-model and parasitic plant phylogeny.

Key words Low copy nuclear genes; Orobanchaceae; Parasites; PPR genes; Phylogeny

Introduction

Parasitic plants are attractive for many scientists and farmers due to their interactions with hosts (e.g. some parasitic plants are serious agricultural pest) and diversity in life styles, which often are correlated with structural and physiological adaptations. Around 1% of angiosperms (includes 22 families, c.4000 species) are parasitic plants (Nickrent et al., 1998), and parasitism has evolved about twelve times within angiosperms (Schneeweiss 2013). Orobanchaceae are the most fascinating family for the evolution research, for its unique characteristic of including the entire set of life styles, ranging from autotrophic nonparasite to facultative/obligate heterotrophic hemiparasite, and lastly to obligate heterotrophic holoparasite. Orobanchaceae is the largest parasitic family in angiosperms. It is of worldwide distribution, especially in (warm) temperate regions, and comprises c. 2060 species in 90 genera (McNeal et al., 2013), in which there are some parasitic weeds, or agricultural pest species parasitic on economically important crop plants (Heide-Jørgensen 2008). In the family, parasitism has evolved once, while the transition from hemi- to holoparasitism has independently evolved multiple times.

As the increasing size of both taxa sampling and molecular data, a certain consensus has been reached to place the nonparasites *Lindenbergia* in Orobanchaceae and to transfer all the hemiparasites that were formerly placed in Scrophulariaceae to Orobanchaceae (Young et al., 1999; Wolfe et al., 2005; Bennett and Mathews 2006), which used to comprise exclusively holoparasitic Orobanchaceae sensu stricto (Beck-Mannagetta 1930). Furthermore, recently Rehmanniaceae (including two nonparasitic genera, *Rehmannia* and *Trianeophora*) was inferred as sister to Orobanchaceae (Xia et al., 2009; Albach et al., 2009), have been transferred to the expanded Orobanchaceae (Angiosperm Phylogeny Group 2016). Several plastid and nuclear loci have been widely used at family level of Orobanchaceae in the previous phylogenetic researches (dePamphilis et al., 1997; Wolfe and dePamphilis 1998; Young et al., 1999; Bennett and Mathews 2006; McNeal et al., 2013). Among these McNeal et al (2013) made enormous progress by including ca 2/3 of the genera, applying plastid and nuclear loci to illustrate phylogenetic relationships within Orobanchaceae. They proved that *PHYA* and *PHYB* are the most useful among all the loci in resolving Orobanchaceae phylogeny, and grouped Orobanchaceae into one nonparasitic clade and five major parasitic clades. Although previous studies have improved our understanding of the phylogenetic relationships in the family, ca 1/3 of the genera of Orobanchaceae remain unplaced (especially in *Buchnereae* clade). Additionally, the relationships among clades in different studies remained either poorly resolved or involved partly well supported incongruences. Firstly, the precise placement of *Brandisia* in Orobanchaceae remains unclear (Bennett and Mathews 2006; McNeal et al 2013; Schneeweiss 2013; Wicke 2013). Secondly, although the position of *Lindenbergia* clade as sister to the rest is well supported by the combined data in McNeal et al (2013), the position of the *Lindenbergia* clade is incongruent between *PHYB* and *PHYA* in McNeal et al (2013), and poorly supported in Bennett and Mathews (2006). Thirdly, the relationships among the *Castilleja-Pedicularis* clade, the *Euphrasia-Rhinanthus* clade and the *Striga-Alectra* clade were contradictory between Bennett and Mathews (2006) and

McNeal et al (2013), despite using the same marker. Lastly, although the phylogenetic placement of clade *Orobanche* were consistent, the monophyly of the holoparasitic *Orobanche* clade was poorly supported by *PHYB*, *ITS* and plastid markers, while better supported by *PHYA* and combined dataset (Bennett and Mathews 2006; McNeal et al., 2013). Part of these controversies may be due to insufficient phylogenetic signal and marker-specific problems, such as evolutionary rate variation of plastid genes (dePamphilis et al., 1997; Schneeweiss 2013; Xi et al., 2015; Wicke et al., 2016). Furthermore, plastid loci are mostly maternally inherited, and may accompany with issues of lower variation rates, aberrant evolution of plastid genomes, relaxed functional constraints on the photosynthetic genes or cytonuclear discordance from introgression (Harris and Ingram, 1991; dePamphilis et al., 1997; Wolfe and dePamphilis 1998; Park et al., 2008; Wicke et al., 2013; Schmickl et al., 2015). ITS is not completely homogenized and may mislead phylogenetic analysis due to paralogy issue (Álvarez and Wendel 2003; Zimmer and Wen. 2012). Although ITS shows no evidence of paralogy issue in Orobanchaceae and its evolutionary rate is high, it has limitation in aligning across family level. Ultimately, *PHYB* and *PHYA* have shortcomings like recent duplicates (paralogous) in Orobanchaceae and problems in amplification. Nowadays PPR and low copy nuclear (LCN) genes have been utilized as tremendous tools in phylogeny of both model and non-model organisms, which may be the ultimate choice in phylogenetic analysis (Sang 2002; Li et al., 2008; Yuan et al., 2009, 2010; Babineau et al., 2013; Crowl et al., 2014).

In earlier study, Zimmer and Wen (2012) have reviewed and summarized a number of LCN loci that can be applied in systematics of various plants, for instance *PHYA* and *PHYB* as mentioned above. Different LCN loci developed from Conserved Ortholog Sets (COS) genes have been tested to be efficient in different angiosperm taxonomic level in various organisms (Li et al., 2008; Duarte et al., 2010; Babineau et al., 2013; Duminil et al., 2015). Moreover, Yuan et al (2009, 2010) have demonstrated that PPR genes can be successfully applied in systematics of model or non-model organisms, at intergeneric or lower taxonomic levels. PPR protein family are sequence-specific RNA-binding proteins, with over 400 members in most eukaryotic plants, although rarely in prokaryotic organism and they are functional in gene expression of chloroplasts and mitochondria (O’Toole et al., 2008; Barkan et al., 2014). 127 PPR genes have been explored as single orthologous in rice (*Oryza sativa*) and *Arabidopsis thaliana*, five of which have been used to resolve the phylogenetic relationships in Verbenaceae with strong molecular evidences (Yuan et al., 2010). Choi et al (2006) tried 274 orthology LCN genes in amplifying 95 species in Fabaceae while only ten markers were evaluated for family level phylogenetic studies. This emphasizes the difficulties in designing proper primers, amplifying and sequencing orthologues LCN loci. Meanwhile, even if we alleviate these obstacles, we still have challenge like paralogs in LCN loci, which should be excluded in plant phylogenetics (Zimmer and Wen. 2012). However, except for two of the phytochrome genes (*PHYA*, *PHYB*), LCN loci have not been applied at family level in Orobanchaceae. Additional data are needed in order to test the hypothesis and resolve the unclear issues in the previous studies, such as to ascertain the position of *Brandisia* and the *Cymbaria-Siphonostegia* clade; the monophyly of the *Orobanche* clade, and the relationships among the *Castilleja-Pedicularis* clade, the *Euphrasia-Rhinanthus* clade and the *Striga-Alectra* clade (Bennett and Mathews 2006; McNeal et al., 2013). Considering that PPR genes and LCN genes are powerful and potentially significant in plant phylogeny, due to their orthologous, unlinked distribution, bi-parentally inherited, rapid evolving and unlinked, universal distribution characteristics in angiosperms, we predict they will be good candidates in resolving phylogenetic problems in Orobanchaceae (Yuan et al., 2009, 2010; Crowl et al., 2014; Zimmer and Wen. 2015). The goals of this study are to utilize PPR and additional LCN genes and try to resolve the uncertain relationships among major clades in Orobanchaceae. In this study, we used two PPR genes previously demonstrated by Yuan et al (2010) as well as three LCN loci in establishing Orobanchaceae phylogeny, and compared our results to previous studies, in order to evaluate their efficiency in resolving intergeneric phylogenetic problems.

Materials and methods

Plant material

We included 54 species out of 29 genera of Orobanchaceae (Table. 1). These taxa covered all major clades identified in previous studies (McNeal et al., 2013; Schneeweiss 2013). Compared to McNeal et al (2013), the most comprehensive phylogenetic study of Orobanchaceae, we have an overall scarcer taxon sampling, especially in the tropical *Striga-Alectra* clade and the *Euphrasia-Pedicularis* clade, but we additionally included *Macrosyringion, Nothobartsia, Odontitella, Phtheirospermum* (except *Phtheirospermum japonicum*), *Pterygiella*, and *Remannia* and *Triaenophora*.

Marker development

Our goal was to establish several low copy markers that amplify well (ideally without requiring any cloning) across the entire family Orobanchaceae. To this end, we tested both already published and newly developed markers. Sequences of the primers and their functions are listed (Table. 2). We tried all the five PPR genes of Yuan et al 2010) and LCN loci, such as degenerate primers *MCM5, MLH1, MSH1* (Zhang et al 2012) and *Agt1, At103, Eif3E, Sqd1, AroB* (Li et al 2008), partly with minor modifications.

In order to retrieve homologous LCN genes from Orobanchaceae, we initially blasted 200+ genes that have been shown to be low copy in *Arabidopsis, Populus, Vitis* and *Oryza* (Duarte et al., 2010) against ESTs from four Orobanchaceae species (*Phelipanche* (*Orobanche*) *aegyptiaca, Triphysaria versicolor, Striga hermonthica, Lindenbergia philippensis*) in Parasitic Plant Genome Project (PPGP, Yang et al, 2014, available at: http://ppgp.huck.psu.edu/blast.php) database to identify homologous LCN genes using tblastx. The cut-off is given as –log(e-value), i.e., 10. The retrieved loci were aligned separately using Muscle 3.8.31, using the web-service available from EMBL-EBI (McWilliam et al 2013). Alignments were edited manually by BioEdit 7.2.1 (Hall 1999). All of the primers were designed in Primer Premier 5.0 software (Premier Biosoft International, Palo Alto, CA) in encoding parts.

Table.1. List of taxa, herbarium, locality and voucher information. NA indicates the value is not available.

Taxon	Herbarium code /Database
Triaenophora shennongjiaensis	China, Hubei, Yichang (HIB: X.D. Li & X. Li Li102)
Rehmannia piasezkii	China, Hubei, Yichang (HIB: X.D. Li & X. Li Li101)
Lindenbergia philippensis	PennState University, USA (cultivated; PAC: S. Wicke LP60/LP61)
Lindenbergia muraria	Material from PSU
Bungea trifida	Turkey, Erzincan (WU: C. Gilli & P. Schönswetter 100-240)
Schwalbea Americana	Newton, GA, USA (PAC: S. Wicke & C. dePampholis SA 57)
Cistanche phelypaea	Spain, Canary Islands, Lanzarote (herb. Sánchez

	Pedraja: G. Moreno Moral MM0038)
Cistanche tubulosa	China, Xinjiang, Cele (HIB: Y.X. Sun Sun201103)
Epifagus virginiana	USA, NH, Dublin (HUH: Boufford et al 42939)
Boschniakia himalaica	China, Yunnan, Yulong snow Mountain (KUN: H.Wang YWB2013057)
Phelipanche arenaria	Austria, Wachau (WU: C. Pachschwöll CP1068)
Phelipanche aegyptiaca	PPGP
Orobanche caryophyllacea	Ukraine, Carpathians (Roman Kish: 2006)
Orobanche flava	Slovakia, Ždiar (WU: C. Pachschwöll CP1069)
Orobanche lycoctoni	Spain, Picos de Europa
Orobanche gracilis	Spain, Palencia, La Pernía (herb. Sánchez Pedraja: G. Moreno Moral MM0110/2007)
Brandisia hancei	China, Yunnan, Songming (KUN: H.Wang YWB2013015)
Rhinanthus alectorolophus	Germany, Baden-Württemberg, Marchtolsheim (MSUN: S. Wicke 5.7.12)
Odontites vernus	SALA 153638 SP
	DP12
Odontites luteus	SALA
	DP1040
Odontites viscosus	SALA 135637 SP
	DP13
Odontites bolligeri	SALA 110065 SP
	DP1036
Odontites cebennensis	SALA 135679 SP
	DP628
Odontitilla virgata	SALA 135636 SP
	DP14
Macrosyringion longiflorum	SALA 135639 SP

	DP11
Bellardia trixago	SALA 142076 SP
	DP918
Nothobartsia asperrima	SALA MAR
Parentucellia latifolia	SALA 142077 SP
	MO6019
Parentucellia viscosa	SALA 142079 SP
	MO6012
Euphrasia stricta	LE568
Euphrasia frigida	Fischer 107
Euphrasia sinuata	Austria, Tirol, Kitzbühler Alpen (WU: D. Pan & G.M. Schneeweiss Schneeweiss120)
Phtheirospermum tenuisectum	China, Yunnan, Zhongdian (HUH: Boufford et al42135)
Pterygiella duclouxii	China, Yunnan, Elephant Mountain (KUN: H.Wang YWB2013166)
Pterygiella cylindrica	China, Yunnan, Daju (KUN: H.Wang YWB2013167)
Melampyrum sylvaticum	Austria, Carinthia, Hohe Tauern (WU: D. Pan & G.M. Schneeweiss119)
Lathraea squamaria	Austria, Vienna, HBV (cultivated)
Triphysaria versicolor	PPGP
Triphysaria pusilla	PPGP
Pedicularis densispica	China, Yunnan, Yulong snow Mountain (KUN: H.Wang YWB2013058)
Pedicularis elwesii	China, Yunnan, Yulong snow Mountain (KUN: H.Wang YWB2013091)
Pedicularis lachnoglossa	China, Yunnan, Yulong snow Mountain (KUN: H.Wang YWB2013092)
Pedicularis rex	China, Yunnan, Fuming old green Mountain (KUN: H.Wang YWB2013154)
Pedicularis verticillata	Austria, Steiermark, Schnaalpe (WU: D. Pan

	Schneeweiss 117)
Pedicularis decora	China, Hubei, Shennongjia (HIB: X.D. Li & X. Li Li106)
Pedicularis aspleniifolia	Austria, Steiermark, Schladminger Tauern (WU: D. Pan & G.M. Schneeweiss Schneeweiss 107)
Pedicularis rostrato-spicata	Austria, Steiermark, Schladminger Tauern (WU: D. Pan & G.M. Schneeweiss Schneeweiss 97)
Radamaea montana	E.Fischer 10292 (?)
Buchnera hispida	E.Fischer10295 (?) LE 176
Buchnera americana	AL,2013 (MSUN)
Striga bilabiata	E.Fischer Le 466
Striga gesnerioides	USA, Florida, Lake County (FLAS 220552: S.F. Brockington 380)
Striga hermonthica	PPGP
Aeginetia indica	China, Jiangxi, Lushan (HIB: Y.S. Peng Peng201201)

Table.2. Primer sequences used in this study for the amplification

Primer		Reference
Name	Sequence	
AT1G04780		
AT1G04780f	CMCTTCATYTGGCTGTTA	This study
AT1G04780r	TCYGDCGAGTCCATYTTA	This study
AT1G04780r511	GGAGMACCWGCACCATCCAA	This study

AT1G14610		
AT1G14610f	RAGGCTAGARGAKGGDAACT	This study
AT1G14610r	AAACTGCCACCAYGARTA	This study
AT1G09680		
AT1G09680_180f	ACCRCCCTWTCTCAAGCCATCCAAA	Yuan et al.,2010
AT1G09680_1760r	TARTCAAGAACAAGCCCTTTCGCAC	Yuan et al.,2010
AT1G09680_850f	GTTAGTTTCAATACTTTGATGAA	Yuan et al.,2010
AT1G09680_850r	TTCATCAAAGTATTGAAACTAAC	Yuan et al.,2010
AT2G37230		
AT2G37230_320f	GCCTGGACDACMCGTTTRCAGAA	Yuan et al.,2010
AT2G37230_1770r	TCRAACAAGCTCTCCATCAC	Yuan et al.,2010
AT2G37230_1800r	GCYGTCTGAACWCSYCCATCYTC	Yuan et al.,2010
AT2G37230_512f	GGCAACAARGTYGAGTAAG	Yuan et al.,2010
AT2G37230_1066r	GATGAGGATTTGTGGGT	This study
Agt1		
Agt1f_oro	GATTTCCGCATGGAYGARTGGGG	Modified from Li et al.,2008
Agt1r_oro	CCAYTCCTCCTTCTGASTGCAGTT	Modified from Li et al.,2008

DNA extraction, PCR and sequencing

Total genomic DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. We have amplified five PPR genes and 90 LCN genes based on different primers. Nevertheless, most of them could only be amplified and sequenced in limited scope. 27 of the 90 primer pairs (30%) gave reliable PCR amplification from three to 23 species across the family (Appendix Table1), but failed with broader taxonmy. We finally successfully amplified five loci gave reliable PCR amplification from at least 37 species at family level in Orobanchaceae, which are LCN gene *Agt1* using modified forward and reverse primers from Li et al (2008), two LCN genes (AT1G04780, AT1G14610) from our study and two PPR genes (AT1G09680 and AT2G37230) using primers from Yuan et al 2010 (Table. 2). We found that for AT1G09680, 180f and 1760r primers were successfully used in amplification, and 850f and 850r primers in sequencing. As for AT2G37230, 320f and 1770r were most successfully used in amplification, and 512f and 1066r primers as secondly successfully. AT1G04780f and AT1G04780r were almost equally successfully used in amplification comparing to AT1G04780f and AT1G04780r511 were most successfully used in amplification. Most loci were directly sequenced without cloning, except a few species in *Buchnereae* clade in AT1G14610 and AT2G37230. Amplification was done in a volume of 15.8 µL containing 0.3 U of KAPA3G Plant DNA Polymerase (Peqlab, Vienna, Austria), 7 µL of 2× PCR buffer, 0.5 µL of 10 µM primers, 0.7 µL DNA and 7 µL water. PCR conditions for LCN loci amplification were: denaturation for 4 min at 94 °C; 35 cycles each with 30 s at 94 °C, 30 s at 48 °C, 1 min at 72 °C; and final elongation for 10 min at 72 °C. For the PPR loci we used the protocol of Yuan et al. (2010). Furthermore, we newly generated some sequences in *PHYA, PHYB, matk* and *rps2* and the PCR conditions are performed as Li et al (2016). PCR products were purified using 0.5 µL Exonuclease I and 1 µL FastAP thermo sensitive alkaline phosphatase (Fisher Scientific, St. Leon-Rot, Germany) following the protocol or, for the cloned samples, with PCR Purification Kit (Qiagen, Germany). 5 µL of purified template, 2 µL trehalose and 1.5 µL sequencing buffer 0.5 µL of primer (10µM) and 1 µL BigDye Terminator were used in cycle sequencing (Applied Biosystems, Foster City, California), cleaned with Sephadex G-50 Fine (GE Healthcare Bio-Sciences, Uppsala, Sweden) and sequenced on an ABI 3730 DNA Analyzer capillary sequencer (Applied Biosystems). For cloning, purified PCR products were run on an agarose gel and target bands were isolated using Quick Gel Extraction Kit (Invitrogen, USA). All PCR products were ligated to vector pGEM-T (Zoman, Beijing) and then were transformed into DH5alpha. After blue white screening on LB medium, eight white colonies were checked by colony PCR, and at least three positive colonies were sequenced with primers M13F and M13R in BGI (Wuhan, China).

Phylogenetic analyses

Sequences were assembled and edited using SeqMan II 5.05 (DNAStar Inc., Madison, USA). Initial alignment of single loci were made with Muscle 3.8.31 (Edgar 2004) using the web-service available from EMBL-EBI (McWilliam et al 2013) and manually adjusted using BioEdit 7.2.1 (Hall 1999). These five loci were combined (the matrix containing 54 species) in order to be more informative, and also test the resolution power of them. Furthermore, we generated a concatenated alignment with 34 species by combining five loci in this study with five loci from McNeal et al. (2013). The best-fit substitution models as well as partitioning schemes for DNA sequence alignments were identified via the Akaike Information Criterion (AIC; Akaike 1974) using PartitionFinder 1.1.0 (Lanfear 2012) employing the greedy algorithm. We only tested those 24 models that are implemented in MrBayes. Maximum likelihood analyses were conducted using RAxML 8.1 (Stamatakis et al., 2014) employing the fast bootstrap approach (Stamatakis et al., 2008) with 2,000 bootstrap replicates. Bayesian inference was done using MrBayes 3.2.3 (Ronquist et al., 2003). Values for all parameters, such as the shape of the gamma distribution or the substitution rates, were estimated during the analysis. Partitions were allowed to evolve under different rates (ratepr = variable). We ran four cold Monte Carlo Markov (MCMC) chains simultaneously starting from different random starting trees for 20 million generations, and sampled trees every 20,000th generation. We used Tracer v.1.4 (Rambaut and Drummond 2007) to check the stability of output parameters from Bayesian analyses. After combining 900 trees from each run (i.e., after discarding 10 % trees as burn-in, when the MCMC chain had reached stationarity, evident from standard deviations of split variances being below 0.01), posterior probabilities were estimated and were plotted on a majority rule consensus tree.

Results

Sequence characteristics and amplification

The five markers were successfully amplified from most of the 54 taxa (Table. 3). Their lengths ranged from 261 bp in *Agt1* to 1461 bp in AT1G09680, and the two PPR genes comprised longer sequences (Table. 4). Introns were present in AT1G14610 and *Agt1*. The intron from *Agt1* was excluded from phylogenetic analysis because it was unalignable across the entire family (Table. 4).

Table.3. Amplification efficiency, GenBank and PPGP information.

Taxon	AT1G04780	AT1G14610	*Agt1*	AT2G37230	AT1G09680
Triaenophora shennongjiaensis	+	+	+	+	+
Rehmannia piasezkii	+	+	+	+	+
Lindenbergia philippensis	+	+	+	+	+
Lindenbergia muraria	+	+			+
Bungea trifida	+	+	+	+	+
Schwalbea americana	+	+	+	+	+
Cistanche phelypaea	+	+		+	+
Cistanche tubulosa	+	+		+	+
Epifagus virginiana	+	+	+	+	
Boschniakia himalaica	+			+	+
Phelipanche arenaria	+	+			
Phelipanche aegyptiaca[1] (OrAeBC4)	3282	38149	47726	18904	487068
Orobanche caryophyllacea	+	+		+	+
Orobanche flava	+	+	+	+	+
Orobanche lycoctoni	+	+		+	+
Orobanche gracilis	+	+		+	+
Brandisia hancei	+	+	+	+	+
Rhinanthus alectorolophus[Li]	+	+	AM503660	+	+
Odontites vernus	+		+	+	+
Odontites luteus	+		+	+	+

Odontites viscosus	+	+		+	+
Odontites bolligeri	+	+	+	+	+
Odontites cebennensis	+	+		+	+
Odontitilla virgata	+	+	+	+	+
Macrosyringion longiflorum	+	+	+	+	
Bellardia trixago		+		+	+
Nothobartsia asperrima	+	+	+	+	+
Parentucellia latifolia	+	+	+	+	
Parentucellia viscosa				+	
Euphrasia stricta		+		+	
Euphrasia frigida	+	+	+	+	
Euphrasia sinuata		+	+	+	+
Phtheirospermum tenuisectum	+	+		+	
Pterygiella duclouxii	+	+	+	+	+
Pterygiella cylindrica	+	+	+	+	+
Melampyrum sylvaticum[Li]		+	AM503643	+	+
Lathraea squamaria[Li]	+	+	AM503659	+	+
Triphysaria versicolor[2]	248550	16026	77562	26606	252996
Triphysaria pusilla[3]	8008		880		14880
Pedicularis densispica	+	+	+	+	+
Pedicularis elwesii	+	+	+	+	+
Pedicularis lachnoglossa	+	+	+	+	+
Pedicularis rex		+			+
Pedicularis verticillata	+		+	+	+

Pedicularis decora	+		+	+	+
Pedicularis aspleniifolia	+	+	+	+	+
Pedicularis rostrato-spicata	+	+	+	+	+
Radamaea montana	+		+	+	+
Buchnera hispida	+	+			
Buchnera americana	+	+	+	+	
Striga bilabiata	+	+	+	+	
Striga gesnerioides	+	+	+	+	
Striga hermonthica[4]	299603	31411	408	298151	291302
Aeginetia indica	+			+	
Total number	48	45	37	49	41

+ marked the sequences that newly obtained in this study

1-4 marked the sequences from EST library or combined bilds in PPGP database. 1:OrAeBC4; 2: TrVeBC2; 3: TrPuRnBC1; 4: StHeBC1.The numbers indicated the EST numbers. E.g. OrAeBC4_3282

Li marked the sequences downloaded from GenBank (Li et al., 2008).

Total number marked the numbers of sequences that available in different loci

Table.4. Sequences characteristics

Locus	**Sequence length(bp)**	**Alignment length(bp)**
	Exon（Intron）	**Exon（Intron）**
AT1G04780	435-807	807
AT1G14610 *	250-384 (55-97)	387（**104**）
AT1G09680	786-1458	1461
AT2G37230	391-1359	1359
*Agt1_oro**	206-415（220-818）	261

* loci that contain an intron; numbers in the bracket indicate the intron length.

Phylogenetic analyses

Phylogenetic framework with different markers

The relationships among the major clades are illustrated (Figs. 1-6). We separated the loci we used here into three categories, depending on their abilities of resolution. The best markers turned out to be the two PPR genes: AT1G09680 as the longest locus provided greater resolution than the other four loci, and it showed good resolution at backbone, genera and species levels in all clades (Fig. 1). Another PPR gene AT2G37230 as the second longest locus showed less support at backbone, but good among genera and species levels in all clades, except *Euphrasia-Rhinantheae* clade (Fig. 2). Nevertheless, they are some conflicts between the PPR genes concerning the *Cymbaria-Siphonostegia* clade and *Brandisia*. The second category contained LCN loci AT1G14610 (Fig. 3) and AT1G04780 (Fig. 4), which have never been tested in other studies, showed bad resolution at backbone, but showed partial resolution among genera and species in some clades, e.g. the *Castilleja-Pedicularis* clade, the *Euphrasia-Rhinanthus* clade and the *Striga-Alectra* clade. The third category was *Agt1* (Fig. 5) as the shortest LCN locus provided extremely bad resolution at all levels in all clades, except the *Striga-Alectra* clade.

In this study, we have extensions compared to previous studies. Firstly, the phylogenetic placement of *Triaenophora and Rehmannia* was strongly supported by two PPR genes (Fig. 1, 2) and AT1G14610 (Fig. 3). The unresolved relationships of *Triaenophora and Rehmannia* in *Agt1* (Fig. 5) were poorly supported (BS=43, PP=0.64). *Rehmannia piasezkii* nested within the six major clades in AT1G04780 (Fig. 4), but this was equally poorly supported (BS=54, PP=0.91). Secondly, the newly added *Phtheirospermum* species and *Pterygiella* formed a separate clade (*Pterygiella* clade) with strong support in all five loci. Lastly, we confirmed the placement of the newly added genera *Macrosyringion*, *Nothobartsia*, *Odontitella* in the *Euphrasia-Rhinanthus* clade. It was supported by two PPR genes (Fig. 1, 2) and that *Nothobartsia* and *Odontitella* are sister taxa.

Consistent with previous studies, the monophyly of parasites in Orobanchaceae was well supported by AT1G09680 (Fig. 1), *AT1G14610* (Fig. 3), modestly supported by AT2G37230 (Fig. 2) and AT1G04780 (Fig. 4) and without support from *Agt1*. The monophyly of each of the six clades (e.g. nonparasitic *Lindenbergia* clade, hemiparasitic *Cymbaria-Siphonostegia*, *Castilleja-Pedicularis*, and *Pterygiella* clades, the mostly hemiparasitic *Euphrasia-Rhinanthus* and *Striga-Alectra* clades and the holoparasitic *Orobanche* clade were well supported by AT1G09680 and AT2G37230. The monophyly of the *Lindenbergia* clade, the

Castilleja-Pedicularis clade and the *Striga-Alectra* clade were modestly to well supported by AT1G14610 and AT1G04780, while in *Agt1* (Fig. 5), only the monophyly of the *Striga-Alectra* clade was strongly supported.

Nevertheless, there are discrepancies with the results of McNeal et al (2013).
1) The *Orobanche* clade was highly supported to be monophyletic in AT1G09680 (Fig. 1) and AT2G37230 (Fig. 2), and it is located at the base of the whole parasites with strong support in AT1G09680 and less supported in AT2G37230.
2) The *Euphrasia-Rhinanthus* clade and the *Striga-Alectra* clade as sister was well supported by AT1G09680, and less supported by AT2G37230.
3) The *Pterygiella* clade as sister to clade comprising the *Euphrasia-Rhinanthus* clade and the *Striga-Alectra* clade inferred by AT1G09680.
4) The *Cymbaria-Siphonostegia* clade and *Brandisia* were at the base or nested within the mostly hemi-parasites clades. However, the exact positions of the *Cymbaria-Siphonostegia* clade and *Brandisia* remained contradicting between AT1G09680 and AT2G37230. The placement of *Brandisia* was unresolved in three LCN genes, and had been poorly supported by AT2G37230 (BS=55, PP=97, Fig. 2) as sister to the clade comprising the *Castilleja-Pedicularis* clade, the *Euphrasia-Rhinanthus* clade, the *Pterygiella* clade and the *Striga-Alectra* clade, while on the other hand it was sister to the *Castilleja-Pedicularis* clade supported by AT1G09680 (BS=85, PP=0.99, Fig. 1).

Phylogenetic framework with combined datasets

Facing the controversies above we combined all the five markers in our study. Combined dataset showed a robust phylogenetic structure and good resolution among genera and species, but with insufficient resolution at part of the backbone, possibly due to the conflicts of the PPR genes that have been mentioned above. The combined data set comprised 4379 nucleotide sites in 54 species (Fig. 6). The *Lindenbergia* clade was sister to the united of the parasites clades. The *Orobanche* clade, that comprises only holoparasitic species was highly supported to be monophyletic (BS=94, PP=1) and as sister to the other parasites with strong support (BS=100, PP=1). *Brandisia* as sister to the mostly hemiparasitic clades (including *Cymbaria-Siphonostegia* clade) was well supported (BS=96, PP=1). The *Euphrasia-Rhinanthus* clade was inferred as sister to the *Striga-Alectra* clade, with moderate support (BP=84, PP=1). *Pterygiella* was sister to *Phtheirospermum* (except *Phtheirospermum japonicum*), and they formed a new clade, the *Pterygiella* clade, with high support (BP=100, PP=1). *Nothobartsia* was highly supported as sister to *Odontitella*, and *Macrosyringion* as sister to monophyletic *Odontites* was well supported (BP=100, PP=1).

Nevertheless, we still have some challenges in the combined dataset:

1) The *Castilleja-Pedicularis* clade as sister to the clade (*Cymbaria-Siphonostegia* clade (*Pterygiella* clade (*Euphrasia-Rhinanthus* clade, *Striga-Alectra* clade))) was poorly supported (BP=58, PP=0.99, Fig. 6)

2) The position of the *Cymbaria-Siphonostegia* clade was unresolved, but was nested within the hemi-parasites.

3) Monophyly of *Odontites* was only weakly supported (BP=58, PP=0.90, Fig. 6).

In order to better illustrate the issues above, we combined our loci with the five loci from McNeal et al (2013), and this matrix comprised 11165 nucleotide sites in 34 species (Fig. 7). The placement of *Triaenophora*, *Rehmannia*, the *Lindenbergia* clade and the *Orobanche* clade remained unchanged. However, slightly unlike the result from combined data of five loci in our study (Fig.6), *Brandisia* was sister to the clade comprising the *Castilleja-Pedicularis* clade, the *Euphrasia-Rhinanthus* clade, the *Pterygiella* clade and the *Striga-Alectra* clade, but without *Cymbaria-Siphonostegia* clade (BS=92, PP=1, Fig. 7). Oppositely, combined data of ten loci inferred that the *Striga-Alectra* clade was sister to the *Castilleja-Pedicularis* clade, with strong support (BP=100, PP=1, Fig. 7), but not as sister to the *Euphrasia-Rhinanthus* clade as inferred from combined data of five loci (BP=84, PP=1, Fig. 6). The *Pterygiella* clade was again well supported. However, different from the result from combined data of five loci in our study (Fig.6), *Pterygiella* clade as sister to the *Euphrasia-Rhinanthus* clade was strongly supported and they were together sister to the clade comprising the *Castilleja-Pedicularis* clade and the *Striga-Alectra* clade. But the relationships among the *Cymbaria-Siphonostegia* clade and the *Orobanche* clade and the rest remained unclear.

Discussion

Availability and effectiveness of sequence data from PPR and three LCN loci

We have tested five new LCN loci in 29 genera (54 species) in phylogeny of Orobanchaceae, in order to reconstruct the relationships among the major clades and also to evaluate the effectiveness of these new LCN loci compared to the previous studies. As illustrated (Table. 4), PPR genes, AT1G14610 and AT1G04780 can be more easily amplified, while for *Agt1*, although being the shortest marker, only 37 sequences could be obtained out of 54 species.

We confirm the high potential of PPR genes for molecular phylogenetic studies at family or lower levels (Yuan et al. 2010; Crowl et al., 2014). PPR genes evolve rapidly and play important roles in translation and expression in chloroplasts and mitochondria (O’Toole et al., 2008). They are distributed universally in eukaryotes, and most of the PPR genes are single orthologous and intronless (Yuan et al., 2009; Barkan et al., 2014).) Furthermore, as indicated by Crowl et al (2014), PPR genes are able to track past hybridization events and may be helpful in inferring phylogeny in species level. These imply that conservative PPR proteins are powerful phylogenetic markers (Yuan et al 2009, 2010; Crowl et al. 2014). The PPR genes AT1G09680 and AT2G37230 are not only readily amplifiable, straightforward in alignment, but also much longer compared to the other LCN loci (Tables. 3, 4). They are more informative in resolving intergeneric and interspecific relationships (Figs. 1, 2). Although AT2G37230 has been indicated to have recent gene duplication in genera *Glandularia* and *Verbena* (Yuan et al., 2010), it appears to have no paralogues in our study (Fig. 2). AT1G09680 and AT2G37230 show generally congruent relationships among the major six clades, except for the ambiguous position of *Brandisia* and the *Cymbaria-Siphonostegia* clade (Figs. 1, 2).

In order to resolve the ambiguous position of *Brandisia* and the *Cymbaria-Siphonostegia* clade, we need to use multiple, independent LCN loci besides PPR genes. Hence, we tested and developed some LCN loci that readily amplified and aligned (Table. 3). AT1G14610 and AT1G04780 code for an Aminoacyl-tRNA ligase and and Ankyrin repeat family protein, respectively (Ometto et al 2012; Isner et al 2012), and have never been used in phylogenetic studies till now. They are shorter in length and with fewer informative sites than the two PPR genes (Table. 4). On the other hand, as described by Liepman and Olsen (2003), *Agt1* encodes a peroxisomal photorespiratory enzyme involved in photorespiration. Although it has been corroborated useful in phylogenetic reconstruction (Li et al 2008; López-Pujol 2012; Gonzalez 2014), it contains less informative signals at family level here when compared with the other four loci (Table. 4, Fig. 5).

Overall, the topologies for Orobanchaceae in PPR genes (Figs. 1, 2) and the combined dataset (Figs. 6, 7) are better supported comparing to the other three LCN loci (Figs. 3-5), and some conflicting topologies resulted from poorly supported values. The PPR genes used in our study furthermore confirmed the outstanding achievement of PPR proteins in non-model and parasitic plant phylogeny, although they have not been widely used in large scape of plant phylogeny studies till now. As our expectation, the relationships among the major clades were poorly supported by the three LCN genes, which may result from their shorter sequences length (Table. 4). Intergeneric relationships are almost resolved by LCN loci, but they have limited use at family level, especial for *Agt1*.

The limitation of PPR genes and LCN loci is the difficulties in designing proper primers, amplifying and sequencing orthologues LCN loci, for their low copy characteristics and various evolutionary histories (Crowl et al., 2014). Some loci may succeed in certain taxonomical level, but not at a broad range (Li et al., 2008). So we could only find a few loci that worked for our purposes, although we tried many. For example, the PPR genes (except AT1G09680 and AT2G37230) that worked in Yuan et al (2010) but failed to work in this study may due to loci differentiation between Verbenaceae and Orobanchaceae, which might due to genome reduction, evolutionary rate variation and pseudogene formation in Orobanchaceae.

Phylogenetic relationships among the major clades

Although there are minor conflicts among different marker in some major clades, the phylogenetic relationships of major clades identified in PPR markers (Fig. 1, 2) are mostly consistent with combined data (Fig. 6, 7). Relationships remain unresolved in combined data due to contradicting results from single markers. The monophyly of parasites and each of the six clades and the relationships within the major clades are strongly supported by combined dataset (Fig. 6, 7) as in McNeal et al (2013). While the holoparasitism has independently evolved at least three times has been corroborated in this study again (Fig. 6, 7), the first is *Orobanche* clade,which contains exclusively holoparasites; the second is *Lathraea* which is the only holoparasite in the *Euphrasia-Rhinanthus* clade, and the last is *Aginetia* , which is one of the holoparasites in the *Striga-Alectra* clade.

We included two newly sampled autotrophic species *Triaenophora shennongjiaensis* and *Rehmannia piasezkii* (Table. 1). *Triaenophora*. (6 species) and *Rehmannia* (3 species) are endemic to China (Chin, 1979; Li et al., 2005; Li et al., 2008). We regarded them as outgroup here.

The *Lindenbergia* clade includes only *Lindenberia* (12 species), which is an Afro-Asiatic autotrophic genus (Hjertson et al., 1995). Although the position of the *Lindenbergia* clade is incongruent between *PHYA* and other four markers in McNeal et al (2013), we infer this clade as the basal clade within the six major clades of Orobanchaceae, and it is sister to the united parasitic clades which receives strong support in either single marker (except AT1G04780 (Fig. 4), and *Agt1* (Fig. 5) or combined dataset (Fig. 6, 7). This is also well supported by the other loci in McNeal et al (2013), by plastid loci (Young et al 1999; Olmstead et al 2001), and by pollen morphological characters (Hjertson et al., 1995; Bennett and Mathews 2006).

Against most previous suggestions, the *Orobanche* clade is strongly supported as sister to all other parasitic clades (including the *Cymbaria-Siphonostegia* clade). The *Orobanche* clade contains around 10 genera (ca.180 species), and all are exclusively non-photosynthetic holoparasites, which used to be regarded as Orobanchaceae sensu stricto (Beck 1930). Our phylogenetic structure indicates that the holoparasitic *Orobanche* clade has been separated from the united hemiparasitic clades. While, this is contrast with McNeal et al (2013) and Bennett and Mathews (2006), which placed the holoparasitic *Orobanche* clade nested within the mostly hemiparasitic clades.

Brandisia as a distinct lineage is corroborated, but its precise position could not be ascertained in the previous studies. *Brandisia* is nested at the base of the mostly hemiparasitic clades except the *Cymbaria-Siphonostegia* clade highly supported by ten loci combined data (BS=92, PP=1, Fig. 7). Although we receive higher support in the placement of *Brandisia* by combined dataset than previous studies (McNeal et al 2013; Bennett and Mathews 2006), it behaves as a rogue taxon, with contradicting positions in different PPR genes, well supported in AT1G09680 (Fig. 1), and less supported in AT2G37230 (BS=55, PP=0.97, Fig. 2).

The *Cymbaria-Siphonostegia* clade *Cymbarieae* contains five hemiparasitic genera (ca.20 species) distributed mainly in Eurasia, but also with disjunct distribution (McNeal et al., 2013). The mostly hemiparasitic clades *Cymbarieae D.Don, Castillejeae,* and *Rhinantheae s.s* together were regarded as tribe Rhinantheae s.l. by Fischer (2004) based on molecular evidences, but the tribe was supposed to be polyphyletic, due to the unclear phylogenetic position of the *Cymbaria-Siphonostegia* clade. The *Cymbaria-Siphonostegia* clade as the first split at the base in parasites was poorly resolved by single nuclear marker in some studies (Bennett and Mathews 2006; McNeal et al 2013), but was well supported by combined data in McNeal et al 2013. While, the placement of the *Cymbaria-Siphonostegia* clade is unclear in both combined data (Figs. 6, 7) in this study, due to its contradicting positions in the two PPR genes. But we present evidence that the *Cymbaria-Siphonostegia* clade is nested within or closely related to the mostly hemiparasitic clades, which are well supported by AT1G09680 (Fig. 1), and AT2G37230 (Fig. 2), and in both combined data, which is consist with the delimitation of tribe Rhinantheae s.l. by Fischer (2004).

We included additional species from the debated *Pterygiella* complex taxa: *Phtheirospermum tenuisectum*, *Pterygiella duclouxii* and *Pterygiella cylindrica,* which are endemic to southwestern China (Table. 1). Circumscription of *Pterygiella* clade is corroborated by ten loci combined data (Fig. 7) assister to *Euphrasia-Rhinanthus* clade in the previous studies, although the *Striga-Alectra* clade is not included in Dong et al 2013 (Bennett and Mathews 2006; Dong et al 2013; McNeal et al 2013). While on the other hand, morphological evidences indicated that three genera *Xizangia, Phtheirospermum* and *Pterygiella* from the *Pterygiella* clade were closely related to *Brandisia* and *Lindenbergia*, due to the similar surface of fruit and reticulate seeds (Dong et al 2015). It is controversial whether the *Pterygiella* clade is sister of the clade comprising the *Euphrasia-Rhinanthus* clade and the *Striga-Alectra* clade as indicated here AT1G09680 (Fig. 1) and five loci combined data (Fig. 6) or close to the *Euphrasia-Rhinanthus* clade as indicated here in ten loci combined data (Fig. 7) and also by McNeal et al (2013) and Dong et al (2013). More molecular and morphology data are needed in order to resolve the precise position of the *Pterygiella* clade.

Lastly, the relationships among the *Castilleja-Pedicularis* clade, the *Euphrasia-Rhinanthus* clade, and the *Striga-Alectra* clade are not well resolved, even if the *Cymbaria-Siphonostegia* clade is ignored. There two hypotheses concerning the relationships among the above three clades. One is from Bennett and Mathews 2006, who regarded that the *Euphrasia-Rhinanthus* clade is sister to the *Striga-Alectra* clade (MPBS=99, MLBS=82) with *PHYA*, taking the paralogs from three species in the *Euphrasia-Rhinanthus* clade into account, and they together are sister to the *Castilleja-Pedicularis* clade (MPBS=100, MLBS=100). Another one is from McNeal et al (2013), who considered that the *Castilleja-Pedicularis* clade is sister to the *Striga-Alectra* clade (MLBS=99, with *PHYA*, excluding paralogs of the three species; MLBS=100, with both *PHYB* and combined dataset), and then they are strongly supported as sister to the *Euphrasia-Rhinanthus* clade in *PHYA*, *PHYB*, and combined data. The PPR genes and five loci combined dataset in our study support the hypothesis from Bennett and Mathews (2006) and regarded that the *Euphrasia-Rhinanthus* clade and the *Striga-Alectra* clade are sister taxa with well supported in five loci combined dataset (Fig. 6), AT1G09680 (Fig. 1) and poorly supported in AT2G37230 (MLBS=71, PP=0.97, Fig. 2). However, this contradicted the hypothesis of McNeal et al (2013) and the ten loci combined dataset (MLBS=100, PP=1, Fig. 7) here. In the future, we need more PPR or LCN genes or genomic data to resolve such complex phylogenetic problems.

Conclusions

Although it took us plenty of effort to test different PPR and LCN genes, we confirmed that PPR genes are well suited to address questions of parasitic plant phylogeny in family or lower levels, while the single LCN markers show less insufficient resolving power. Here, major clades previously identified are confirmed. Against most previous suggestions, the holoparasitic *Orobanche* clade is consistently supported as sister to all parasitic clades, including the hemiparasitic *Cymbaria-Siphonostegia* clade. The newly added *Phtheirospermum* species and *Pterygiella* form a separate clade (*Pterygiella* clade). Unclear relationships do, however, remain (e.g., the precise placement of *Brandisia* and relationships among the *Cymbaria-Siphonostegia* clade, the *Castilleja-Pedicularis* clade, the *Euphrasia-Rhinanthus* clade, and the *Striga-Alectra* clade. In the future, more markers are needed through phylogenomic approach to resolve the unclear relationships, when single markers are still insufficient.

Acknowledgements We thank Da Pan for sampling and help with the software Patitionfinder and Clemence Pachschwöll, Xiaodong Li, Susann Wicke, Yanxia Sun and all the friends who help us with sampling or providing material. Special thank goes to Sarah Mathews (Australian National University) and Elfriede Grasserbauer and Michael H.J.Barfuss (University of Vienna) for their help with lab work. Financial support from the program of China Scholarships Council (No.201206100007) to X. L. is gratefully acknowledged.

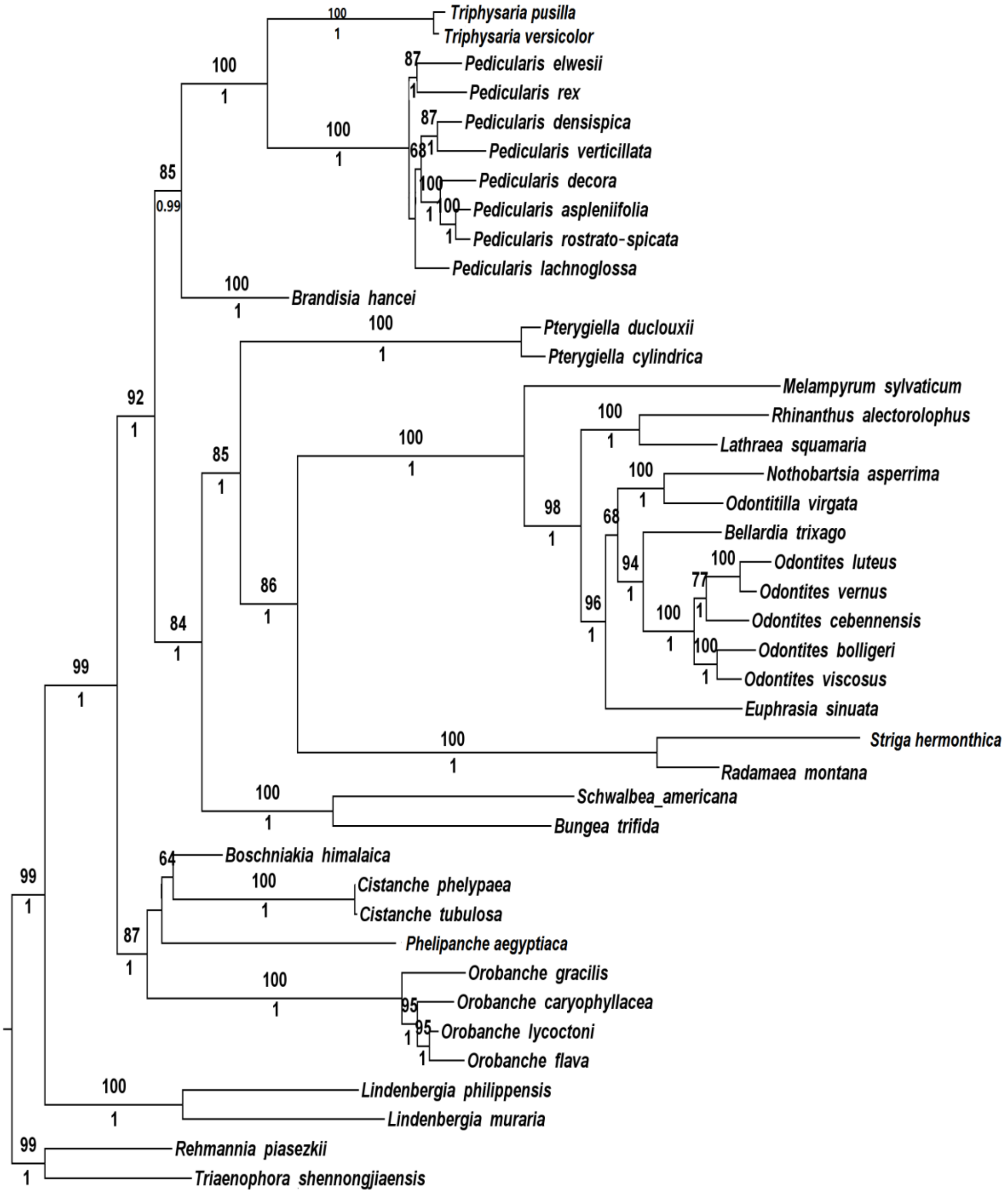


Fig.1. Phylogenetic relationships within Orobanchaceae inferred using maximum likelihood on an AT1G09680 data set. Numbers at branches are maximum likelihood bootstrap support values (50 or higher) and posterior probabilities (0.95 or higher). *Rehmannia piasezkii* and *Triaenophora shennongjiaensis* were chosen as outgroup

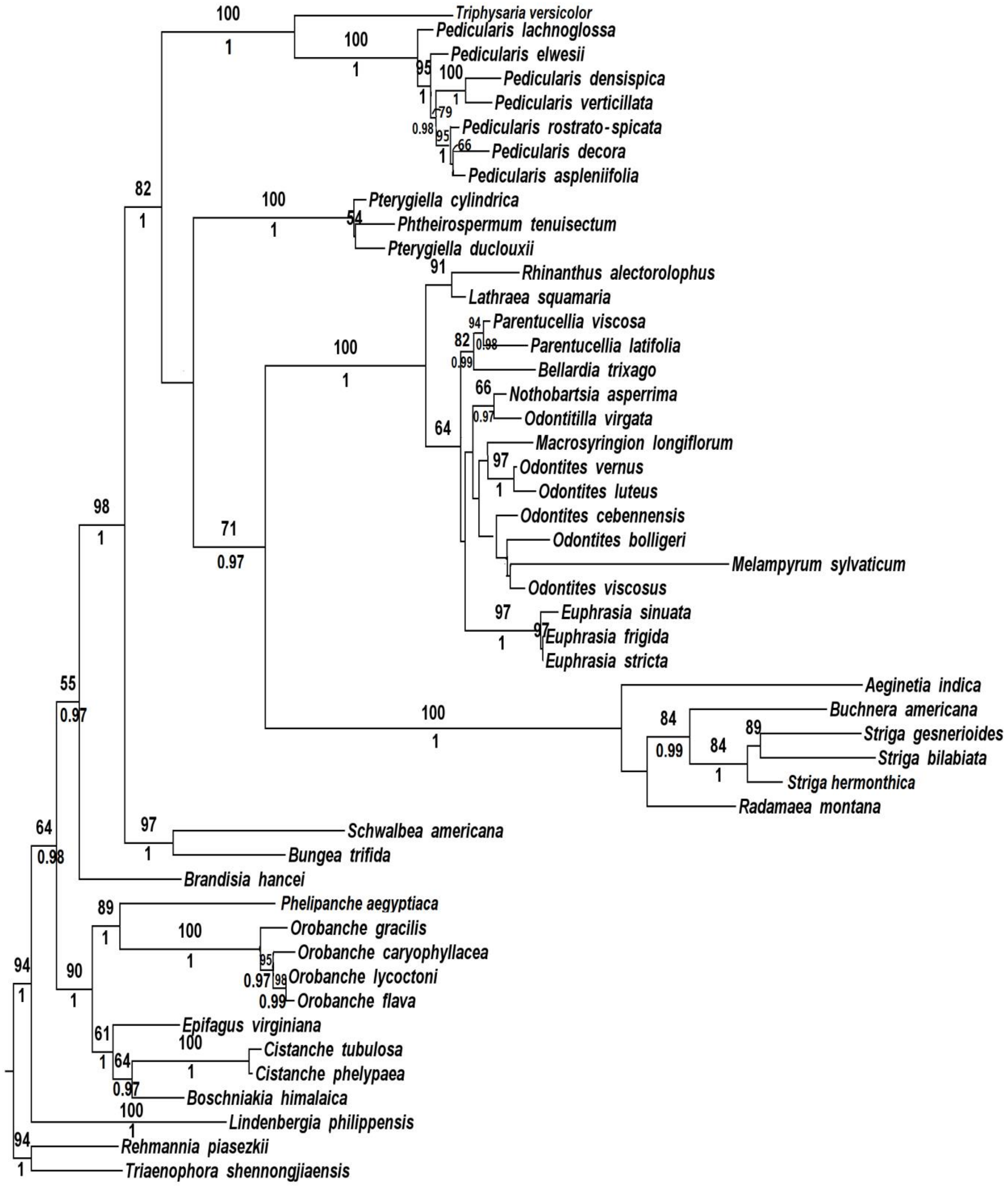


Fig. 2. Phylogenetic relationships within Orobanchaceae inferred using maximum likelihood on an AT2G37230 data set. Numbers at branches are maximum likelihood bootstrap support values (50 or higher) and posterior probabilities (0.95 or higher). *Rehmannia piasezkii* and *Triaenophora shennongjiaensis* were chosen as outgroup.

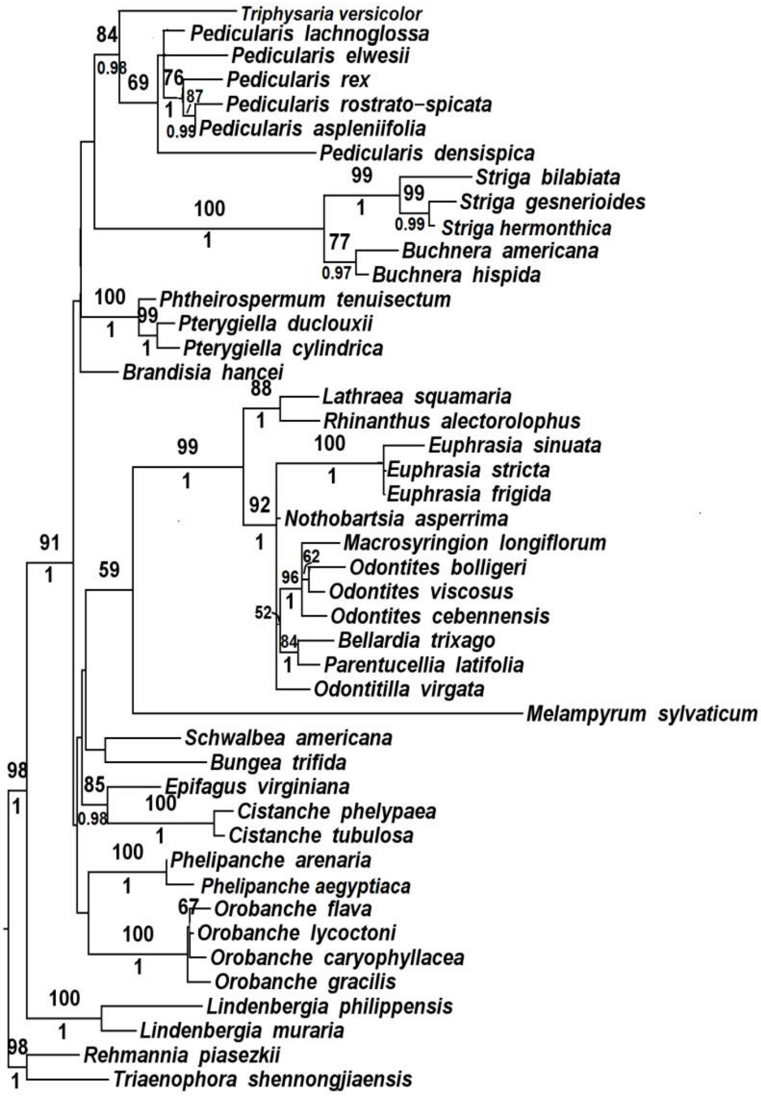


Fig. 3. Phylogenetic relationships within Orobanchaceae inferred using maximum likelihood on an AT1G14610 data set. Numbers at branches are maximum likelihood bootstrap support values (50 or higher) and posterior probabilities (0.95 or higher). *Rehmannia piasezkii* and *Triaenophora shennongjiaensis* were chosen as outgroup.

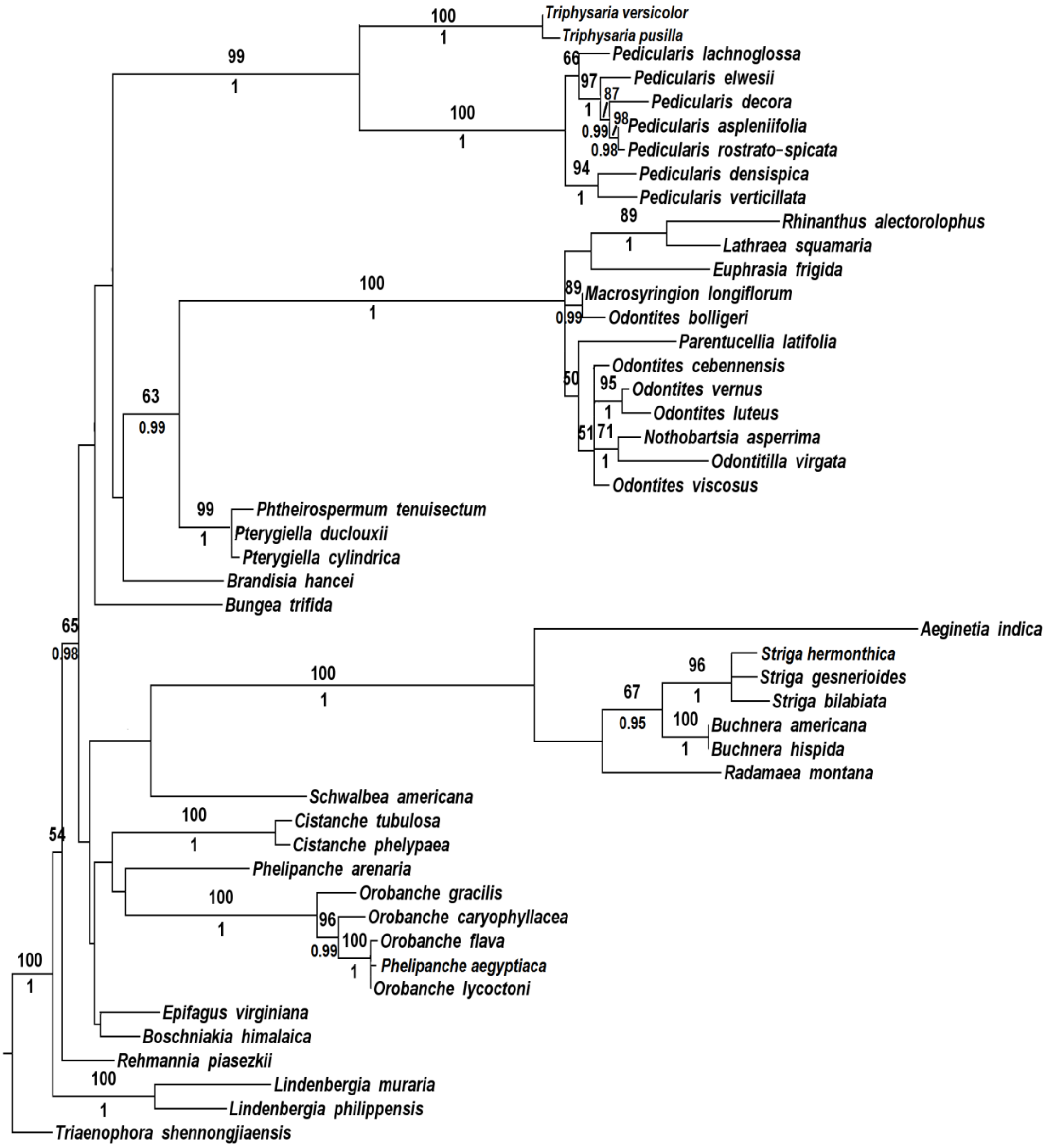


Fig.4. Phylogenetic relationships within Orobanchaceae inferred using maximum likelihood on an AT1G04780 data set. Numbers at branches are maximum likelihood bootstrap support values (50 or higher) and posterior probabilities (0.95 or higher). *Triaenophora shennongjiaensis* was chosen as outgroup.

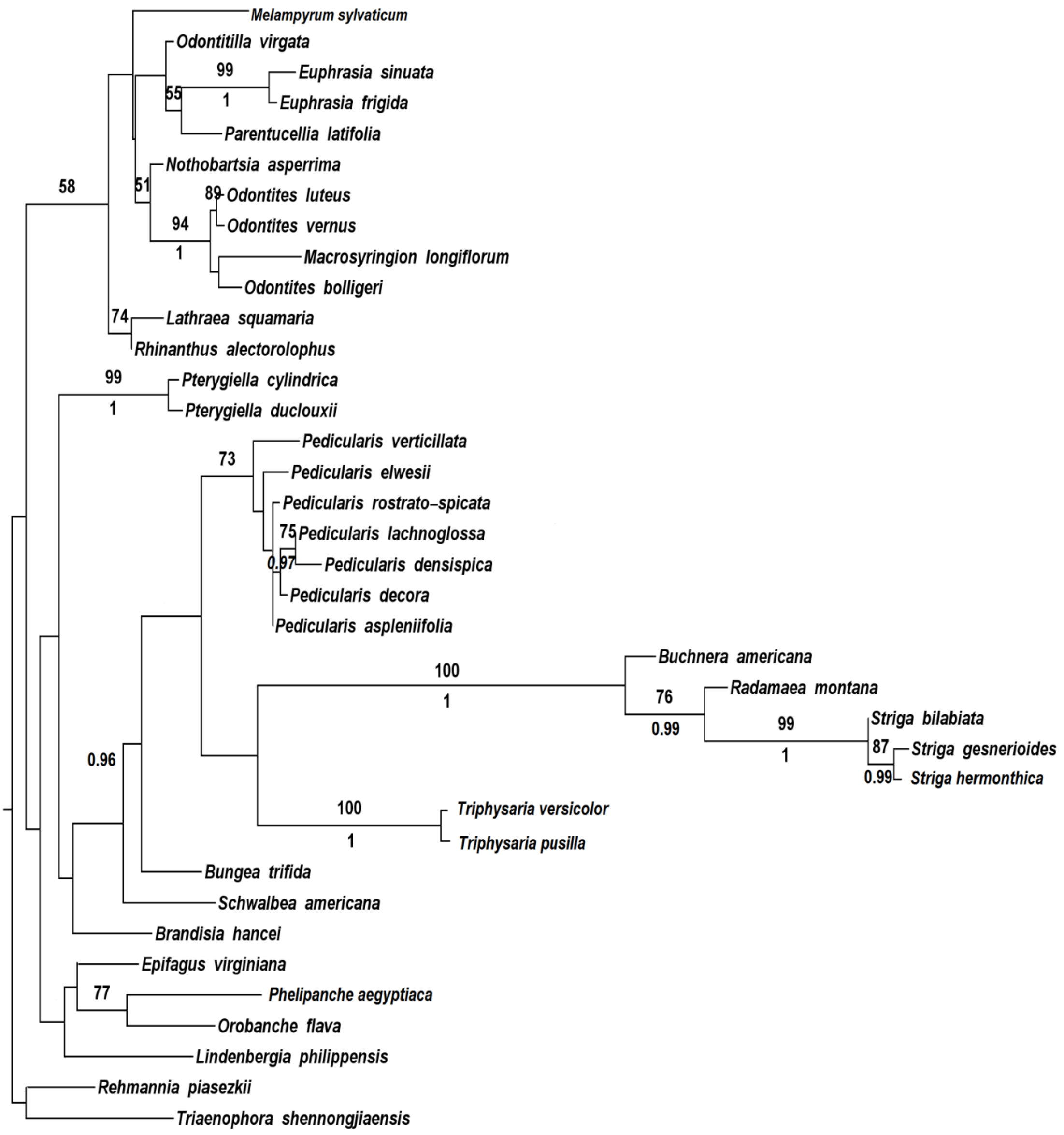


Fig. 5. Phylogenetic relationships within Orobanchaceae inferred using maximum likelihood on an *Agt1* data set. Numbers at branches are maximum likelihood bootstrap support values (50 or higher) and posterior probabilities (0.95 or higher). *Rehmannia piasezkii* and *Triaenophora shennongjiaensis* were chosen as outgroup.

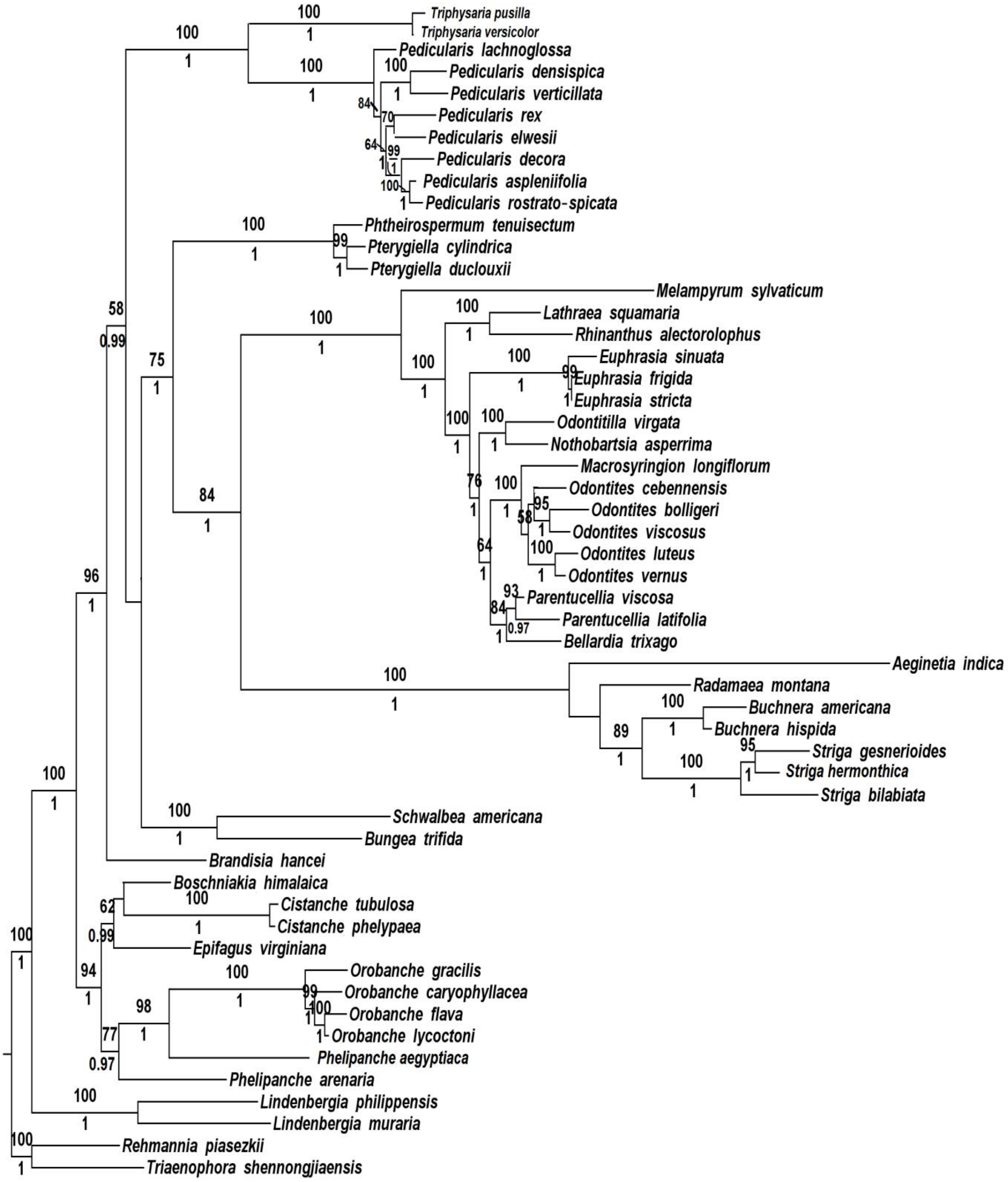


Fig. 6. Phylogenetic relationships within Orobanchaceae inferred using maximum likelihood on combined data set. Numbers at branches are maximum likelihood bootstrap support values (50 or higher) and posterior probabilities (0.95 or higher). *Rehmannia piasezkii* and *Triaenophora shennongjiaensis* were chosen as outgroup.

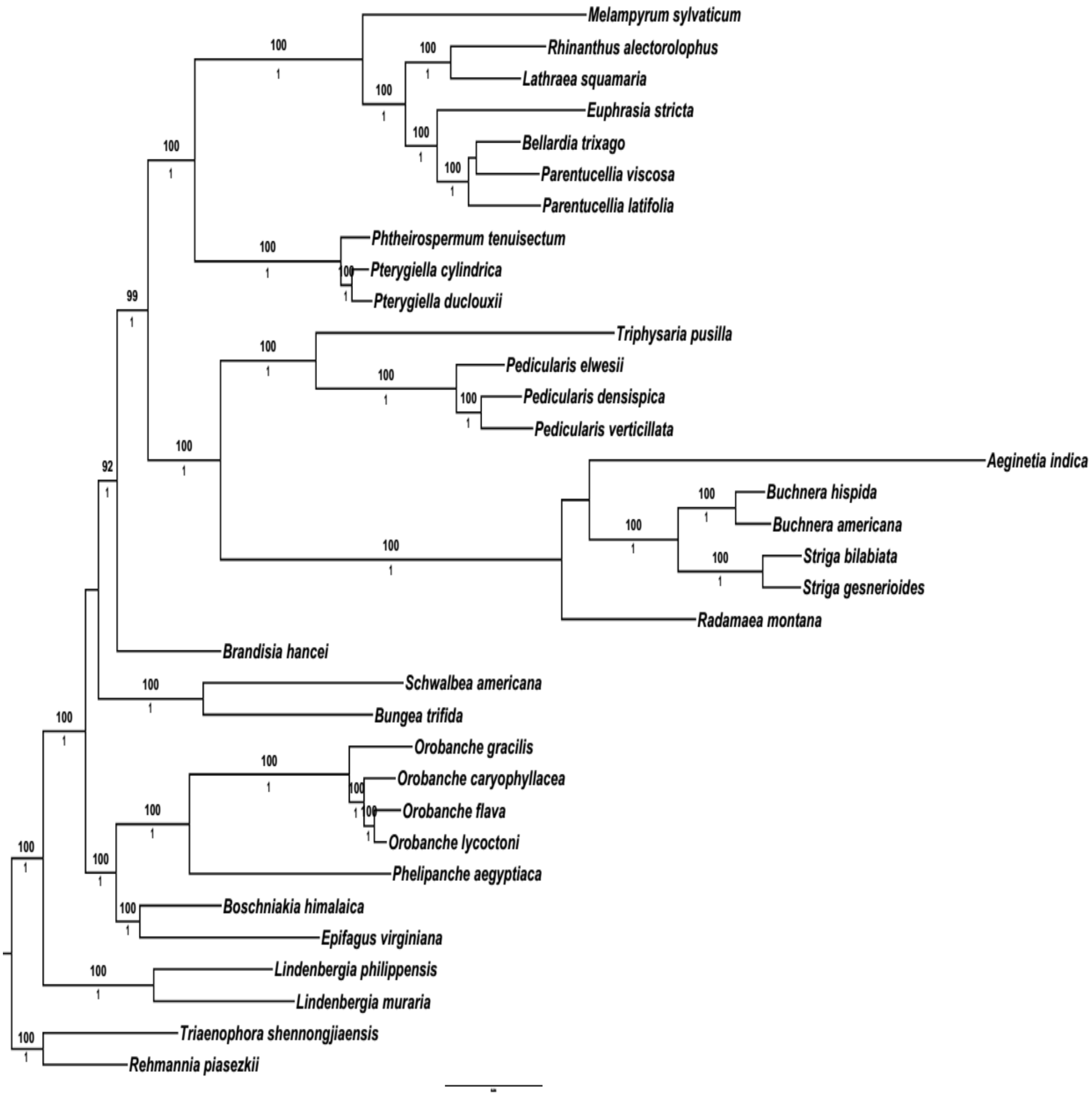


Fig. 7. Phylogenetic relationships within Orobanchaceae inferred using maximum likelihood on combined data set of ten loci, five from this study and five from Mcneal's study. Numbers at branches are maximum likelihood bootstrap support values (90 or higher) and posterior probabilities (0.99 or higher). *Rehmannia piasezkii* and *Triaenophora shennongjiaensis* were chosen as outgroup.

References

Angiosperm Phylogeny Group (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society, 181 (1): 1–20, doi:10.1111/boj.12385

Albach DC, Yan K, Jensen, SR, Li H-Q (2009) Phylogenetic placement of *Triaenophora* (formerly Scrophulariaceae) with some implications for the phylogeny of Lamiales. Taxon 58:749–756

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716-723.

Álvarez I, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. Mol Phylogen Evol 29:417–434. doi: 10.1016/S1055-7903(03)00208-2

Babineau M, Gagnon E et al (2013) Phylogenetic utility of 19 low copy nuclear genes in closely related genera and species of caesalpinioid legumes. South African Journal of Botany 89: 94-105.

Barkan A, Small I (2014) Pentatricopeptide repeat proteins in plants. Annu. Rev. Plant Biol., 65, pp. 415–442

Bennett J, Mathews S (2006) Phylogeny of the parasitic plant family Orobanchaceae inferred from phytochrome A. Amer J Bot 93:1039–1051

Beck-Mannagetta G (1930) Orobanchaceae. In: Engler A (ed) Das Pflanzenreich IV 261. Wilhelm Engelmann, Leipzig, pp 1–348

Chin T. L (1979) *Rehmannia* and *Triaenophora*. InP. C. Tsoong and H. P. Yang [eds.], Flora reipublicae popularis sinicae, vol. 67, part 2, Scrophulariaceae, 212 – 222. Science Press, Beijing, China.

Choi, H.K., Luckow, M., Doyle, J., Cook, D.R (2006) Development of nuclear gene-derived molecular markers linked to legume genetic maps. Molecular Genetics and Genomics 276, 56–70

Crowl AA, Mavrodiev E, Mansion G, Haberle R, Pistarino A, Kamari G, et al. (2014) Phylogeny of Campanuloideae (Campanulaceae) with Emphasis on the Utility of Nuclear Pentatricopeptide Repeat (PPR) Genes. Louis EJ, editor. PLoS ONE;9: e94199. doi: 10.1371/journal.pone.0094199.s022. pmid:24718519 dePamphilis CW, Young ND, Wolfe AD (1997) Evolution of plastid gene *rps2* in a lineage of hemiparasitic and holoparasitic plants: many losses of photosynthesis and complex patterns of rate variation. Proc Natl Acad Sci USA 94:7367–7372

Duarte J.M, Wall P.K, Edger P.P, Landherr L.L, Ma H, Pires J.C, Leebens-Mack J, dePamphilis C.W(2010) Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryzaand* their phylogenetic utility across various taxonomical levels. BMC Evolutionary Biology 10, 61–79

Duminil J. et al. (2015) Late Pleistocene molecular dating of past population fragmentation and demographic changes in African rain forest tree species supports the forest refuge hypothesis. J. Biogeogr., 42(8), 1443-1454.

Dong LN, Wang H, Wortley AH, Lu L, Li DZ (2013) Phylogenetic relationships in the *Pterygiella* complex (Orobanchaceae) inferred from molecular and morphological evidence. Bot J Linn Soc 171:491–507

Dong, L.-N., H. Wang, et al. (2015). Fruit and seed morphology in some representative genera of tribe Rhinantheae sensu lato (Orobanchaceae) and related taxa. Plant Systematics and Evolution 301(1): 479-500.

Edgar R. C (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32.

Fischer E (2004) Scrophulariaceae. In: Kadereit JW, ed.The families and genera of vascular plants, Vol. VII. Berlin:Springer Press, 333–432.

Gonzalez L.A (2014) Phylogenetics and mating system evolution in the southern South American *Valeriana* (Valerianaceae). M.S. Thesis, Department of Biological Sciences, University of New Orleans, Louisiana, U.S.A.

Harris S.A and Ingram R (1991) Chloroplast DNA and Biosystematics: The Effects of Intraspecific Diversity and Plastid Transmission. Taxon 40(3): 393-412.

Hall T.A (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for 402 Windows 95/98/NT, in: Nucleic Acids Symposium Series. 95–98.

Heide-Jørgensen HS (2008) Parasitic flowering plants. Brill, Leiden

Hjertson M. L (1995) Taxonomy, phylogeny, and biogeography of *Lindenbergia* (Scrophulariaceae). Botanical Journal of the Linnean Society 119: 265 – 321

Isner J.C, Nuhse T, Maathuis F.J (2012) The cyclic nucleotide cGMP is involved in plant hormone signalling and alters phosphorylation of *Arabidopsis thaliana* root proteins. J. Exp. Bot , 63, 3199–3205.

Lanfear R, Calcott B, et al (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution, 29, 1695–1701.

Li M, Wunder J, Bissoli G, Scarponi E, Gazzani S, Barbaro E, Saedler H, Varotto C (2008) Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species. Cladistics 24,727–745.

Li, X., T.-S. Jang, et al (2016) Molecular and karyological data confirm that the enigmatic genus *Platypholis* from Bonin-Islands (SE Japan) is phylogenetically nested within *Orobanche* (Orobanchaceae). Journal of Plant Research: 1-8.

Li X.D, Li J.Q, and Zan Y.Y (2005) A new species of *Triaenophora* (Scrophulariaceae) from China. Novon 15: 559 – 561.

Li X. D, Zan Y. Y, Li J.Q, and Yang S. Z (2008) A numerical taxonomy of the genera *Rehmannia* and *Triaenophora* (Scrophulariaceae). Journal of Systematics and Evolution 46: 730 –737.

Liepman A.H, Olsen L J (2003) Alanine aminotransferase homologs catalyze the glutamate: glyoxylate aminotransferase reaction in peroxisomes of Arabidopsis. Plant Physiol. 131, 215–227.

López-Pujol J, Garcia-Jacas N, Susanna A, Vilatersana R (2012) Should we conserve pure species or hybrid species? Delimiting hybridization and introgression in the Iberian endemic Centaurea podospermifolia. Biological Conservation 152:271–279.

McNeal J, Bennett JR, Wolfe AD, Mathews S (2013) Phylogeny and origins of holoparasitism in Orobanchaceae. Am J Bot 100(5): 971-983

McWilliam H. et al. (2013) Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res. 41.

Nickrent D. L, Duff R. J et al (1998) Molecular Phylogenetic and Evolutionary Studies of Parasitic Plants. Molecular Systematics of Plants II: DNA Sequencing. D. E. Soltis, P. S. Soltis and J. J. Doyle. Boston, MA, Springer US: 211-241.

Nickrent D.L., Musselman L.J(2004) Introduction to parasitic flowering plants. Plant Health Instructor. doi:10.1094/PHI-I-2004-0330-01.

O'Toole N, Hattori M, Andr´ es C, Iida K, Lurin C, et al (2008) On the expansion of the pentatricopeptide repeat gene family in plants.Mol. Biol. Evol.25:1120–28

Park JM, Manen JF, Colwell AE, Schneeweiss GM (2008) A plastid gene phylogeny of the non-photosynthetic parasitic *Orobanche* (Orobanchaceae) and related genera. J Plant Res 121:365–376

Ometto L, Li M, Bresadola L, Varotto C (2012) Rates of evolution in stress-related genes areassociated to habitat preference in two *Cardamine* lineages. BMC Evolutionary Biology, *12*, *7*.

Olmstead R.G, dePamphilis C.W, Wolfe A.D, Young N.D, Elison W.J and Reeves P.A (2001) Disintegration of the Scrophulariaceae. Amer. J. Bot.88: 348-362.

Rambaut A., Drummond A.J (2007) Tracer v1.4. Available from: <http:// tree.bio.ed.ac.uk/software/tracer/>.

Ronquist F, Huelsenbeck J.P (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574

Sang T (2002) Utility of low-copy nuclear gene sequence in plant phylogenetics. Critical Reviews in Biochemistry and Molecular Biology 37, 121–147.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), 1312–1313. http://doi.org/10.1093/bioinformatics/btu033

Stamatakis A, Hoover P, and Rougemont J. (2008) A rapid bootstrap algorithm for the RAxML web servers. Systematic Biology 57: 758 – 771 .

Schmickl R, Liston A, Zeisek V, Oberlander K, Weitemier K, Straub SCK, et al (2015) Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). Molecular Ecology Resources.In Press doi: 10.1111/1755-0998.12487.

Schneeweiss G.M (2013) Phylogenetic Relationships and Evolutionary Trends in Orobanchaceae. Parasitic Orobanchaceae: Parasitic Mechanisms and Control Strategies. Joel M. D, Gressel J and Musselman J. L. Berlin, Heidelberg, Springer Berlin Heidelberg: 243-265.

Wicke S (2013) Genomic Evolution in Orobanchaceae. Parasitic Orobanchaceae: Parasitic Mechanisms and Control Strategies. Joel M. D, Gressel J and Musselman J. L. Berlin, Heidelberg, Springer Berlin Heidelberg: 267-286.

Wicke, S., K. F. Müller, et al (2016) Mechanistic model of evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants. Proceedings of the National Academy of Sciences 113(32): 9045-9050

Wolfe AD, Randle CP, Liu L, Steiner KE (2005) Phylogeny and biogeography of Orobanchaceae. Folia Geobot 40:115–134

Wolfe AD, and dePamphilis CW (1998) The effect of relaxed functional constraints on the photosynthetic gene rbcL in photosynthetic and nonphotosynthetic parasitic plants. Molecular Biology and Evolution 15: 1243 – 1258 .

Xia Z, Wang YZ, Smith JF (2009) Familial placement and relations of *Rehmannia* and *Triaenophora* (Scrophulariaceae s.l.) inferred from five gene regions. Amer J Bot 96:519–530

Xi Z, Liu L, Davis CC (2015) The impact of missing data on species tree estimation. Mol. Biol. Evol. msv266. (doi:10.1093/molbev/msv266

Yang, Z., Wafula, E. K., Honaas, L. A., Zhang, H., Das, M., Fernandez-Aparicio, M., Huang, K., Bandaranayake, P. C. G., Der, J. P., Clarke, C. R., Ralph, P. E., Landherr, L., Altman, N. S., Timko, M. P., Yoder, J. I., Westwood, J. H., & dePamphilis, C. W. (2014). Comparative transcriptome analyses reveal core parasitism genes and suggest gene duplication and repurposing as sources of structural novelty. Molecular biology and evolution, msu343

Young ND, Steiner KE, dePamphilis CW (1999) The evolution of parasitism in Scrophulariaceae/Orobanchaceae: Plastid gene sequences refute an evolutionary transition series. Ann Miss Bot Garden 86: 876–893

Yuan Y.W., Liu C., Marx H.E., Olmstead R.G (2010) An empirical demonstration of using pentatricopeptide repeat (PPR) genes as plant phylogenetic tools: Phylogeny of Verbenaceae and the *Verbena* complex. Molecular Phylogenetics and Evolution 54: 23 – 35.

Yuan Y.W, Liu C, Marx H.E, Olmstead R.G (2009) The pentatricopeptide repeat (PPR) gene family, a tremendous resource for plant phylogenetic studies. New Phytol. 182, 272–283.

Zhang N, Zeng L, Shan H & Ma H (2012) Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol. 195, 923–937.

Zimmer E. A and Wen J (2012) Using nuclear gene data for plant phylogenetics: Progress and prospects. Molecular Phylogenetics and Evolution 65(2): 774-785.

Zimmer EA, Wen J (2015) Using nuclear gene data for plant phylogenetics: Progress and prospects II. Next-gen approaches. J Syst Evol 53: 371–379. doi: 10.1111/jse.12174

Appendix Table.1. Additional primer sequences used in this study for the amplification

Primer		Reference (Number of amplification)
Name	Sequence	
AT1G12640 717f	GAYTATCTWGAATGGACYGA	This study (7)
AT1G12640 717r	HGATTGAACAAAGAATATGA	This study
AT1G51610 f	GGTTATTCCAAGGAGAGATT	This study (4)
AT1G51610 r	CYGGTCCAATCACYTCGC	This study
AT3G06720 f1	CWGTWTGGGCTTTGGGWAAT	This study (4)
AT3G06720 f2	AYATTGCTTCTGGRACMTCT	This study
AT3G06720 r	TTCTCAGCYTCACCHACCTT	This study
AT3G44110 F	GAYCAGTATGGWGARGATGC	This study (3)
AT3G44110 R	TTRCCCTTCATGAAYGGCCT	This study
AT3G10670 f	YCAKTCKCCAGTTGAAATTC	This study (4)
AT3G10670 r	MGCCTTGTAGCCTTCTYTTT	This study
AT3G20870 f	TTCWTWGAYTTGGCTCAT	This study (4)
AT3G20870 r1	TTCCYTCWGGAAAATTGTG	This study
AT4G02060 R	AATGCTWACAGTTTGCTGCT	This study (23)
AT4G02060 F2	CYGAGGATGGYGATATCTAC	This study
AT4G26900 F	GGTCATTGCYTGTCTTGAT	This study (3)
AT4G26900 R	ACTCKGMMGCAACTTCCAAA	This study
AT5G08420 R	KCACTTGTTTTAGACCTTTA	This study (10)
AT5G08420 F2	AYGATGAGATGCAGTGTGAY	This study
AT5G12370 F	TGCTKCBTGTGAAGAAATGG	This study (10)
AT5G12370 R	GCTTGYTTRTTAAGRCCTTC	This study
AT5G43600 F	ATWCCWAGCAAATCACATCT	This study (9)
AT5G43600 R	GGAATGAAWATCATRCCCAT	This study
AT3G63410 f	TGGACKGAGGAYATGAGRGA	This study (5)
AT3G63410 r	CTGCAAAGGMGARTCYCCAG	This study
AT3G63410 1000r	ACGATACCATTTYGGRCC	This study
AT3G63410 900r	GAAGAGCATCCAMACATCTG	This study
AT3G63410 145f	TATGTTTCGGCWGGRAGGTA	This study
AT3G63410 140f	CGATAGATATGTTTCGGCWG	This study
AT2G33150 F	ATGGCTGCRTTYTATGCT	This study (14)
AT2G33150 R1	TCCWGCACCATCACTYACTT	This study
AT2G33150f1 465	TGTCTYCTYCCRATGGGTGT	This study
AT2G33150r2 542	AGCYTGRTCYTGYTCCTGCC	This study
AT2G33150 f140	AGGCAATGYTCATCTGGC	This study
AT2G33150 r830	CCAGATTTAGGGTCCACAAT	This study

AT2G33150 f95	AAGCTGTAGCTGATGTTGC	This study
AT2G33150 r890	GTGGTTGAGCCATCTTTC	This study
Sqd1f_Li	CTTGGGACSATGGGTGARTATGG	Li et al.,2008 (7)
Sqd1r_Li	CCWACAGCAGCYTGMACACAGAACC	Li et al.,2008
AT3G09060 930F	AGTGCTYTGATTCATGGGTTGTG	Yuan et al.,2010 (4)
AT3G09060 2080R	ACAGCTCKRACAAGTATRTTCCA	Yuan et al.,2010
AT5G39980 550F	CACGGRCTGTTCGACGAAATGCG	Yuan et al.,2010 (7)
AT5G39980 1890R	AGACTCAGCATCTGRAAATGAAC	Yuan et al.,2010

Chapter 3

Marker development for phylogenomics in Orobanchaceae, a family with contrasting nutritional modes

Manuscript has been submitted to BMC Bioinformatics

Marker development for phylogenomics: the case of Orobanchaceae, a plant family with contrasting nutritional modes

Xi Li, Baohai Hao, Da Pan, Gerald M. Schneeweiss*

Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, A-1030 Vienna, Austria

*Correspondence:

Gerald M. Schneeweiss

Department of Botany and Biodiversity Research, University of Vienna

E-mail: gerald.schneeweiss@univie.ac.at

Abstract

Background: Phylogenomic approaches, employing next-generation sequencing (NGS) techniques, have revolutionized systematic and evolutionary biology. Target enrichment is an efficient and cost-effective method in phylogenomics and is becoming increasingly popular. Depending on availability and quality of reference data as well as on biological features of the study system, (semi-) automated identification of suitable markers will require specific bioinformatic pipelines. Here we established a highly flexible bioinformatic pipeline to identify putative orthologous single copy genes (SCGs) and to construct bait sequences for use in the nutritionally heterogeneous plant family Orobanchaceae.

Results: We used transcriptome data of differing quality available for four Orobanchaceae species and, as reference, SCG data from monkeyflower (*Erythranthe guttata,* syn. *Mimulus* *g.*; 1,915 genes) and tomato (*Solanum lycopersicum*; 391 genes). Depending on whether gaps were permitted in initial BLAST searches of the four Orobanchaceae species against the reference, our pipeline identified 1,307 and 981 SCGs respectively, of length of at least 780 bp. Automated bait sequence construction (using 2× tiling) resulted in 38,156 and 21,856 bait sequences, respectively. In comparison to the recently published MarkerMiner 1.0 pipeline ours identified about 1.6 times as many SCGs (of at least 900 bp length).

Conclusions: We developed a new pipeline to obtain multilocus probe sets for Orobanchaceae. This pipeline is also applicable in other non-model plants, including other parasitic plants, where only transcriptomes or partial genome data of differing quality are available.

Keywords: bioinformatic pipeline; marker development; Orobanchaceae; phylogenomics; single copy nuclear genes; target enrichment.

Background

Combining target enrichment with next generation sequencing (NGS) strategies can yield a large number of low copy nuclear (LCN) loci and is becoming increasingly popular for systematic and evolutionary biology [1]. Target enrichment or sequence capture is a DNA-based method that uses hybridization of sheared DNA from the species of interest to conserved oligonucleotides [2]. As it can also be applied to samples with already fragmented DNA as obtained from herbarium specimens, utilizing the wealth of material available in natural history collections becomes feasible [3–4]. Target enrichment relies on the availability of reference data, which may come from draft genomes, transcriptome data (RNASeq or EST data), genome skimming, or a combination thereof [5–7]. For example, Weitemier and colleagues [6] and de Sousa and colleagues [8] used whole genome data to identify 768 genes (≥960 bp) in milkweeds (*Asclepias*) and 50 genes (≥2,000 bp) in burclover (*Medicago*), respectively. Schmickl and colleagues [7] combined genome skimming data from one sorrel species (*Oxalis obtusa*) with transcriptome data from another species (*Oxalis corniculata*) for targeting LCN loci. To this end, the authors implemented an automated and interactive bash script workflow for discovery of LCN loci for phylogenetic analysis. Chamala and colleagues [9] and Mayer and colleagues [10] developed MarkerMiner 1.0 and BaitFisher, respectively. Both tools provide an automated, user-friendly and web-supported workflow for discovery of LCN loci from transcriptome data. Disadvantages of these approaches include reliance on whole or draft genome sequences [6,8] or settings that bias against less conserved loci, i.e., strictly reciprocal blast searches in MarkerMiner [9] and the highly reduced bait-to-target distances in BaitFisher [10]. Additionally, none of these methods assessed the effect of using different blast strategies (with and without gaps) on number and length of recovered LCN loci.

The plant family Orobanchaceae encompass a wide range of nutritional modes from non-parasitic autotrophy via photosynthetic parasitism to non-photosynthetic parasitism and contains some major pest species on economically important crop plants [11]. Orobanchaceae have become a model system for studying the evolution of parasitic plants, including molecular evolution under altered functional constraints [12–18]. Following these research avenues requires a solid phylogenetic framework, but although considerable progress has been made in reconstructing phylogenetic relationships within Orobanchaceae [14–15,19–23], numerous areas of unknown or uncertain relationships remain [24]. Phylogenomic approaches are needed to confidently resolve phylogenetic relationships, but the identification of suitable genomic markers requires specific bioinformatic pipelines. Here we establish a highly flexible bioinformatic pipeline to discover putative orthologous single copy genes (SCGs) from Orobanchaceae and to construct bait sequences. To this end, we used transcriptome data of differing quality available from four Orobanchaceae species and SCG data of monkeyflower

(*Erythranthe guttata* [syn: *Mimulus g.*], Phrymaceae) and tomato (*Solanum lycopersicum*, Solanaceae) as reference dataset.

Implementation

The workflow of the pipeline, which uses publicly available software and tools as well as several python scripts, is shown in Fig. 1. A detailed description of the single steps can be found in Appendix S1. All BLAST searches used BLAST 2.2.6.

The first step is to identify single copy genes in species that are closely related to our group of interest (here Orobanchaceae) and for which well-annotated genomes are available. We chose to use *E. guttata* (http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PhytozomeV10) as primary reference [25] and the more distantly related, but better annotated *S. lycopersicum* (ftp://ftp.solgenomics.net/genomes/Solanum_lycopersicum/annotation/ITAG2.4_release/) as secondary reference species [26]. The putative status as SCG was assessed by comparison against SCG data from *Arabidopsis thaliana* (compiled by Alexander Kozik and Richard Michelmore, available from http://cgpdb.ucdavis.edu/COS_Arabidopsis/data/arabidopsis_single_copy_genes.fasta) using blastx [27]. Briefly, *E. guttata* genes that had sufficiently high similarity to, but only a single match with *Arabidopsis* SCGs were identified and blasted against each other to select genes that are single copy in *Erythranthe*. Using the same strategy, we obtained SCGs from *S. lycopersicum*. Finally, we merged the *Erythranthe* and *Solanum* SCG data sets, retaining only those SCGs from *Solanum* that had no significant blast hit with *Erythranthe* to avoid gene redundancy. The thus obtained reference data sets contained 2,306 SCG loci: 1,915 SCG sequences from *E. guttata*, our primary reference, and 391 SCG loci from *S. lycopersicum*, our secondary reference.

Data for the four Orobanchaceae species (our focal species) were downloaded from the Parasitic Plant Genome Project (PPGP) database (at http://ppgp.huck.psu.edu/download.php [28]): *Lindenbergia philippensis* (LiPhGnB1.fasta), *Triphysaria versicolor* (TrVeBC2.fasta), *Striga hermonthica* (StHeBC2.fasta), *Phelipanche (Orobanche) aegyptiaca* (OrAeBC4.fasta). These were blasted against the reference database of non-redundant SCGs from *E. guttata* and *S. lycopersicum* SCGs. As across broader phylogenetic depths homologues genes might differ in length, we used both ungapped blastx (as done in [5]) as well as gapped blastx searches. Subsequently, we parsed the blast output using tcl_blast_parser_123_V047.tcl (available from http://code.google.com/p/atgc-tools/downloads/detail?name=tcl_blast_parser_123_V047.tcl) and retained as putative single-copy conserved orthologous sequence (COS) the best query hit with identity of at least 40, expectation better than 1e-20 (the cut-off is given as –log(e-value), i.e., 20), and an alignment length of at least 100 positions (on the amino acid level). The minimum length of 100 amino acid positions (i.e., 300 bp) is shorter than the length cut-off of 960 bp used by Weitemier and colleagues [6], who used draft genome and transcriptome data from a congeneric reference species, but longer than that of 150 bp applied by Mandel and colleagues [5]. The combined builds from the parasitic species available from PPGP (all but *Lindenbergia*) may include sequences from their host species, which have to be removed prior to further processing. To this end, we blasted the Orobanchaceae COS sequences against sequences from their putative hosts using blastn and removed those sequences with identity of at least 95% to the host. Host sequences might be from species the parasite has actually been grown on (e.g., *P. aegyptiaca* has been grown on tobacco and *Arabidopsis*) or from close relatives. While this step is specific for parasitic species, it may be applied to non-parasitic species if contamination might be an issue.

In contrast to the approach used by MarkerMiner [9], where putative paralogs are retained, we aimed at removing those as completely as possible. Different unigenes from the same focal taxon (here: any of the four Orobanchaceae) that have been blasted against the same reference single copy gene and that exceed a user-defined overlap cut-off are considered putative paralogues and are removed. The stringency of this approach will be determined both by data quality (e.g., many small unigenes versus fewer longer unigenes) and by the desired number of loci to create baits from.

Sequences from the focal species that have been blasted to the same gene from the reference species were extracted and put together in one folder, named after the reference protein ID. As the quality of the transcriptome data from the four Orobanchaceae species differs considerably, we retained only those loci that contained at least one sequence of *L. philippensis*, which based on the N50 value has the best assembled data among our focal species. Alignment of each SCG was done using MAFFT 7.245 [29] using the E-INS-i method, which is suitable for data sets containing multiple conserved domains and long gaps.

Each aligned SCG was split into smaller alignments each corresponding to an exon, whose boundaries in the genes from the reference species (*Erythranthe* and *Solanum*) have been extracted from the gff3 files [25,30] available from Phytozome 10.3 (http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PhytozomeV10). We have no information on exon-intron boundaries in the four Orobanchaceae species, but assume that gene structure is sufficiently similar to that of *Erythranthe* and *Solanum*. Thus obtained exon sequences were degapped, and exons <120 bp (i.e., the length of the bait sequence) were removed. Probes (bait sequences) were designed automatically (using python scripts) using a length of 120 bp and 2× tiling (i.e., baits overlap by 50%), allowing a maximum overlap of baits of 80 bp (i.e., from a fragment of 160 bp still two baits, overlapping by 80 bp, can be extracted). This threshold was chosen to make better use of the 3’-end of an exon while avoiding unduly high redundancy between baits. Loci that yielded fewer than four baits in *Lindenbergia* (i.e., the best assembled focal species) were removed, retaining loci that had at least 280 bp. To reduce redundancy among baits, probes that shared at least 90% identities among the four species were removed using cd-hit-est (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi). Finally, in order to reduce the chance of targeting plastid genomes, baits that in a blastn search had at least 90% identity with any of the 12 published plastid genomes from Orobanchaceae [16] were removed as well.

Results and Discussion

Gapped versus ungapped blast

For identifying SCGs from the focal species, blastx allowing gaps (gapped blast) or not allowing gaps (ungapped blast) has been used. Except for some minor differences during the sequence alignment step (Appendix S1), their results are processed equally using the same scripts. As expected, using gapped blast resulted in more recovered loci: 2,050 SCGs in gapped blast versus 1,845 SCGs in ungapped blast and, after filtering for presence in *Lindenbergia*, 1,690 in gapped blast versus 1,555 in ungapped blast (a decrease of 17.6% and 15.7%, respectively; Table 1). Although gapped blast recovered more loci, the number of loci shared by all four focal species actually decreased while the number of loci shared by maximally three focal species increased as expected (Fig. 2). The strategy using gapped blast recovered on average longer loci than the one using ungapped blast (Table 1). Therefore, the strategy using gapped blast yielded 38,156 baits (from 1,307 SCGs), which were 1.7 times more than the 21,856 baits (from 981 SCGs) obtained using the strategy with ungapped blast (Table 1).

Table1 Number and characteristics of loci identified from the four species

Species data	SCGs	CDS	Probes
	Number	Number	Number
	(average length)	(average length)	
Gapped blast			
LiPhGnB1	1690 (1186)	5435 (278)	18038
TrVeBC2	1201 (938)	2503 (251)	6235
StHeBC2	1278 (985)	2880 (264)	8328
OrAeBC4	1055 (868)	2097 (259)	5555
Ungapped blast			
LiPhGnB1	1555 (846)	3934 (228)	9312
TrVeBC2	1079 (754)	2094 (220)	3912
StHeBC2	1214 (761)	2396 (222)	4888
OrAeBC4	1031 (739)	1928 (225)	3744

Species data are builds of transcriptome data generated within the Parasitic Plant Genome Project (see http://ppgp.huck.psu.edu for details). The first four letters indicate the species: LiPh, *Lindenbergia philippensis* — TrVe, *Triphysaria versicolor* — StHe, *Striga hermonthica* — OrAe, *Phelipanche (Orobanche) aegyptiaca*.

The choice of blast strategy will depend on the divergence of the studied taxa as well as the desired number of bait sequences. For distantly related taxa (genus-level or above), where indels in alignments of coding sequences are more likely, gapped blast may be necessary to achieve the required number of loci and bait sequences. Ungapped blast, which gives more conservative results both with respect to number and length of recovered loci, may be the preferred choice for closely related taxa (i.e., from the intrageneric to intraspecific level) that may have a history of hybridization or introgression.

Test of MarkerMiner 1.0 with our datasets

We also employed the MarkerMiner 1.0 pipeline [9] to compare its efficacy for developing SCGs with our pipeline. As data quality differs among our focal species, we only considered a single focal species, *L. philippensis*, which has the best quality data (i.e., the longest unigenes). As neither *Erythranthe* nor *Solanum* are included in the reference options available in MarkerMiner 1.0, we created a new reference, congruent with the one used as subject database in our pipeline (*E. guttata* and *S. lycopersicum* SCGs).We ran MarkerMiner 1.0 under the default settings, except for changing the reference. Consistent with MarkerMiner 1.0, which uses gapped blast for reciprocal blast searching, for comparison from our pipeline only the results from the gapped blast strategy were considered.

MarkerMiner 1.0 resulted in 539 loci (≥ 900 bp) compared to 865 loci (≥ 900 bp) recovered by our pipeline, of which 428 loci (≥ 900 bp) were shared by both; the number of shared loci increased to 466, if all 1,690 loci identified by our pipeline (i.e., also including those < 900 bp) were considered (Fig. 3). MarkerMiner identified 73 loci (by design ≥ 900 bp) not recovered by our pipeline. The reduced number of identified loci from MarkerMiner 1.0 likely is due to the default stringency criteria. MarkerMiner employs reciprocal blast, with both tblastn and blastx, while our pipeline uses only blastx. Secondly, in MarkerMiner the default length cut-off is 900 bp (compared to 280 bp in our study) and at least 70% of the query has to be aligned with 70% identity to the subject (compared to no overlap length requirements and 40% identity in our pipeline). Indeed, if the criteria in MarkerMiner 1.0 are relaxed (at least 40% of the query has to be aligned with 40% identity to the subject) in the reciprocal BLAST, the number of recovered loci increases to 1,466 loci. The reason for loci exclusively identified by MarkerMiner might be that this program retains loci with multiple hits from the same focal species, while in our pipeline such a locus would be removed from the focal species (not necessarily from the entire data set). One advantage of our method over Marker Miner may be that it is very flexible in dealing with data, because it is a semi-automated pipeline consisting of several independent custom phython scripts that can be easily modified to meet the demands of different studies.

General comparison with available pipelines

There are also other pipelines available. A major limitation is that both draft genomes [6,8] and genome skimming data [7] are still infrequent, and none of them is available for Orobanchaceae. Another limitation lies in the stringent criteria used by those pipelines [5,9,31], which may result in losing a considerable amount of putatively informative sites by non-recovery of a large number of loci and by reduced length of loci (by not allowing gaps). This might negatively affect species tree estimation, whose accuracy has been found to be positively correlated with the number of putative SCGs, even in the presence of both high incomplete lineage sorting and gene rate heterogeneity [32]. Additionally, these approaches enrich conserved loci, which may negatively affect the power to resolve phylogenetic relationships especially at the species level.

In contrast to our pipeline, MarkerMiner outputs SCG sequences, but not bait sequences. MarkerMiner may, however, be combined with BaitFisher [10], a novel software package specifically for baits design, which requires alignments, for instance as created by MarkerMiner (via MAFFT) as input. BaitFisher has been developed to obtain an optimal set of baits by minimizing the number of baits (by reducing redundancy of baits without gaps or ambiguous nucleotides) while maximizing the number of targeted nucleotide sequences. The disadvantage of BaitFisher is that it does not tolerate any gaps nor ambiguity codes in the start position of a putative bait, rendering it susceptible to low-quality samples.

Conclusion

We established a highly efficient bioinformatic pipeline that employs BLAST searches, alignment, length-filtering, and sorting to discover putative orthologous SCGs from transcriptome data of the plant family Orobanchaceae. Although no (draft) whole genome sequence is available for Orobanchaceae, we were able to develop up to 38,156 baits (representing 1,307 loci) from combined transcriptomes of four Orobanchaceae species. A comparison of our pipeline with MarkerMiner 1.0 suggests that our pipeline recovers 1.6 times as many loci as MarkerMiner (used with its default settings), which ultimately may help in improving the accuracy of species tree estimation. As our pipeline is very flexible and takes into account a number of complications specific to parasitic plants (host sequences) or not (low quality reference data), it should be readily applicable to other non-model plants.

Additional file

Additional file 1: A detailed description of the pipeline.

Abbreviations

COS: conserved orthologous sequence; LCN: low-copy nuclear gene; NGS: next-generation sequencing; PPGP: Parasitic Plant Genome Project; SCG: single copy gene.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Python scripts as well as reference data (SCGs of *Erythranthe* and *Solanum*) are available at Dryad under doi: #####.

Competing interests

The authors declare that they have no competing interests.

Funding

The research was funded by China Scholarships Council grant No. 201206100007 (to X.L.). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

GMS and XL designed the pipeline. BHH wrote the python scripts. DP and XL tested the pipeline. XL and BHH analysed the data. GMS and XL prepared the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Lemmon EM, Lemmon AR. High-throughput genomic data in systematics and phylogenetics. Annu Rev Ecol Evol Syst. 2013;44:99–121.
2. Zimmer EA, Wen J. Using nuclear gene data for plant phylogenetics: Progress and prospects II. Next-gen approaches. J Syst Evol. 2015;53:371–9.
3. Paijmans JLA, Fickel J, Courtiol A, Hofreiter M, Förster DW. Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. Mol Ecol Resour. 2016;16:42–55.
4. Suchan T, Pitteloud C, Gerasimova NS, Kostikova A, Schmid S, Arrigo N, Pajkovic M, Ronikier M, Alvarez N. Hybridization capture using RAD probes (hyRAD), a new

tool for performing genomic analyses on collection specimens. PLoS One. 2016;11:e0151651.

5. Mandel JR, Dikow RB, Funk VA, Masalia RR, Staton SE, Kozik A, Michelmore RW, Rieseberg LH, Burke JM. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. Appl Plant Sci. 2014;2:1300085.

6. Weitemier K, Straub SCK, Cronn RC, Fishbein M, Schmickl R, McDonnell A, Liston A. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. Appl Plant Sci. 2014;2:1400042.

7. Schmickl R, Liston A, Zeisek V, Oberlander K, Weitemier K, Straub SCK, Cronn RC, Dreyer LL, Suda J. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). Mol Ecol Resour. 2015;16,1124–35.

8. de Sousa F, Bertrand YJK, Nylinder S, Oxelman B, Eriksson JS, Pfeil BE. Phylogenetic properties of 50 nuclear loci in *Medicago* (Leguminosae) generated using multiplexed sequence capture and next-generation sequencing. PLoS One. 2014;9,e109704.

9. Chamala S, García N, Godden GT, Krishnakumar V, Jordon-Thaden IE, de Smet R, Barbazuk WB, Soltis DE, Soltis PS. MarkerMiner 1.0: a new application for phylogenetic marker development using angiosperm transcriptomes. Appl Plant Sci. 2015;3,1400115.

10. Mayer C, Sann M, Donath A, Meixner M, Podsiadlowski L, Peters RS, Petersen M, Meusemann K, Liere K, Wägele JW, Misof B, Bleidorn C, Ohl M, Niehuis O. BaitFisher: A software package for multispecies target DNA enrichment probe design. Mol Biol Evol. 2016;33:1875–86.

11. Heide-Jørgensen HS. Parasitic flowering plants. Leiden: Brill; 2008.

12. Westwood JH, Yoder JI, Timko MP, dePamphilis CW. The evolution of parasitism in plants. Trends Plant Sci. 2010;15:227–35.

13. dePamphilis CW, Palmer JD. Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. Nature. 1990;348:337–9.

14. Wolfe AD, dePamphilis CW. The effect of relaxed functional constraints on the photosynthetic gene *rbcL* in photosynthetic and nonphotosynthetic parasitic plants. Mol Biol Evol. 1998;15:1243–58.

15. Young ND, dePamphilis CW. Rate variation in parasitic plants: correlated and uncorrelated patterns among plastid genes of different functions. BMC Evol Biol. 2005;5:16.

16. Wicke S, Müller KF, dePamphilis CW, Quandt D, Wickett NJ, Zhang Y, Renner SS, Schneeweiss GM. Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. Plant Cell. 2013;25,3711–25.

17. Wicke S, Müller KF, dePamphilis CW, Quandt D, Bellot S, Schneeweiss GM. Mechanistic model of evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants. Proc Natl Acad Sci USA. 2016;113:9045–50.

18. Fan W, Zhu A, Kozaczek M, Shah N, Pabón-Mora N, González F, Mower JP. Limited mitogenomic degradation in response to a parasitic lifestyle in Orobanchaceae. Sci Rep. 2016;6:36285.

19. Young ND, Steiner KE, dePamphilis CW. The evolution of parasitism in Scrophulariaceae/Orobanchaceae: Plastid gene sequences refute an evolutionary transition series. Ann Miss Bot Gard. 1999;86:876–93.

20. Bennett JR, Mathews S. Phylogeny of the parasitic plant family Orobanchaceae inferred from phytochrome A. Am J Bot. 2006;93:1039–51.

21. Park JM, Manen JF, Colwell AE, Schneeweiss GM. A plastid gene phylogeny of the non-photosynthetic parasitic *Orobanche* (Orobanchaceae) and related genera. J Plant Res. 2008;121:365–76.

22. McNeal J, Bennett JR, Wolfe AD, Mathews S. Phylogeny and origins of holoparasitism in Orobanchaceae. Am J Bot. 2013;100:971–83.

23. Li X, Jang TS, Temsch EM, Kato H, Takayama K, Schneeweiss GM. Molecular and karyological data confirm that the enigmatic genus *Platypholis* from Bonin-Islands (SE Japan) is phylogenetically nested within *Orobanche* (Orobanchaceae). J Plant Res. In press; doi:10.1007/s10265-016-0888-y.

24. Schneeweiss GM. Phylogenetic relationships and evolutionary trends in Orobanchaceae. In: Joel DM, Gressel J, Musselman LJ, editors. Parasitic Orobanchaceae. Wien & al.: Springer; 2013. p. 243–65.

25. Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, Schmutz J, Willis JH, Rokhsar DS. Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. Proc Natl Acad Sci USA. 2013;110:19478–82.

26. Fernandez-Pozo N, Menda N, Edwards JD, Saha S, Tecle IY, Strickler SR, Bombarely A, Fisher-York T, Pujar A, Foerster H, Yan A, Mueller LA. The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. Nucl Acids Res. 2015;43(D1):D1036–41.

27. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

28. Yang Z, Wafula EK, Honaas LA, Zhang H, Das M, Fernandez-Aparicio M, Huang K, Bandaranayake PCG, Wu B, Der JP, Clarke CR, Ralph PE, Landherr L, Altman NS, Timko MP, Yoder JI, Westwood JH, dePamphilis CW. Comparative transcriptome analyses reveal core parasitism genes and suggest gene duplication and repurposing as sources of structural novelty. Mol Biol Evol. 2015;32:767–90.

29. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl Acids Res. 2002;30:3059–66.

30. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012;485:635–41.

31. Mandel JR, Dikow RB, Funk VA. Using phylogenomics to resolve mega-families: an example from Compositae. J Syst Evol. 2015;53:391–402.

32. Xi Z, Liu L, Davis CC. The impact of missing data on species tree estimation. Mol Biol Evol. 2016;33:838–60.

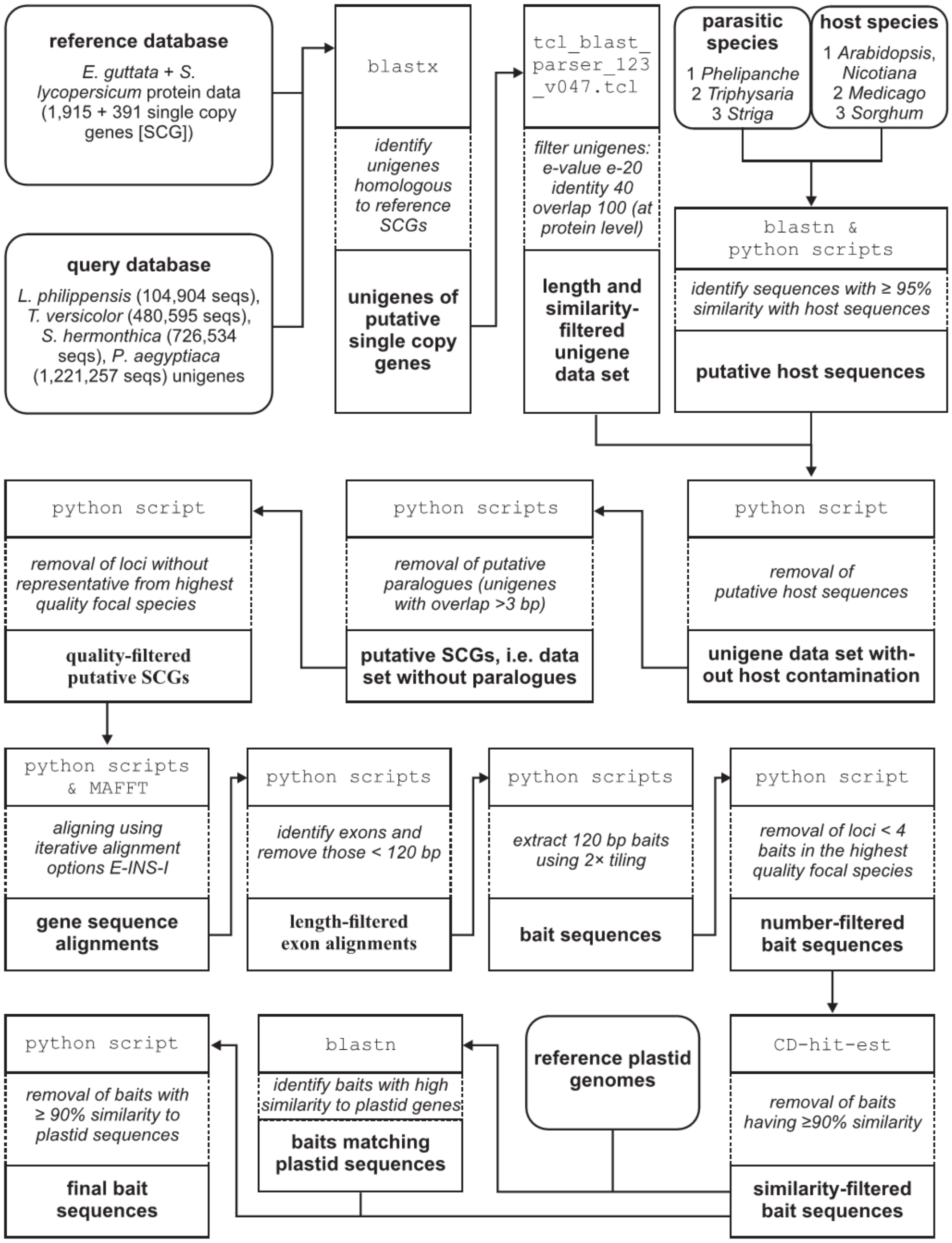

reference database
E. guttata + S. lycopersicum protein data (1,915 + 391 single copy genes [SCG])
query database
L. philippensis (104,904 seqs), T. versicolor (480,595 seqs), S. hermonthica (726,534 seqs), P. aegyptiaca (1,221,257 seqs) unigenes
blastx
identify unigenes homologous to reference SCGs
unigenes of putative single copy genes
tcl_blast_parser_123_v047.tcl
filter unigenes: e-value e-20 identity 40 overlap 100 (at protein level)
length and similarity-filtered unigene data set
parasitic species
1 Phelipanche
2 Triphysaria
3 Striga
host species
1 Arabidopsis, Nicotiana
2 Medicago
3 Sorghum
blastn & python scripts
identify sequences with ≥ 95% similarity with host sequences
putative host sequences
python script
removal of putative host sequences
unigene data set without host contamination
python scripts
removal of putative paralogues (unigenes with overlap >3 bp)
putative SCGs, i.e. data set without paralogues
python script
removal of loci without representative from highest quality focal species
quality-filtered putative SCGs
python scripts & MAFFT
aligning using iterative alignment options E-INS-I
gene sequence alignments
python scripts
identify exons and remove those < 120 bp
length-filtered exon alignments
python scripts
extract 120 bp baits using 2× tiling
bait sequences
python script
removal of loci < 4 baits in the highest quality focal species
number-filtered bait sequences
CD-hit-est
removal of baits having ≥90% similarity
similarity-filtered bait sequences
reference plastid genomes
blastn
identify baits with high similarity to plastid genes
baits matching plastid sequences
python script
removal of baits with ≥ 90% similarity to plastid sequences
final bait sequences

Fig. 1 Workflow of the pipeline. External data are indicated by rounded boxes, data generated within the pipeline by thick-outlined boxes; programs and/or scripts are indicated by `Courier New` font in thin-outlined boxes; the rationale of each step is given in dashed outlined boxes.

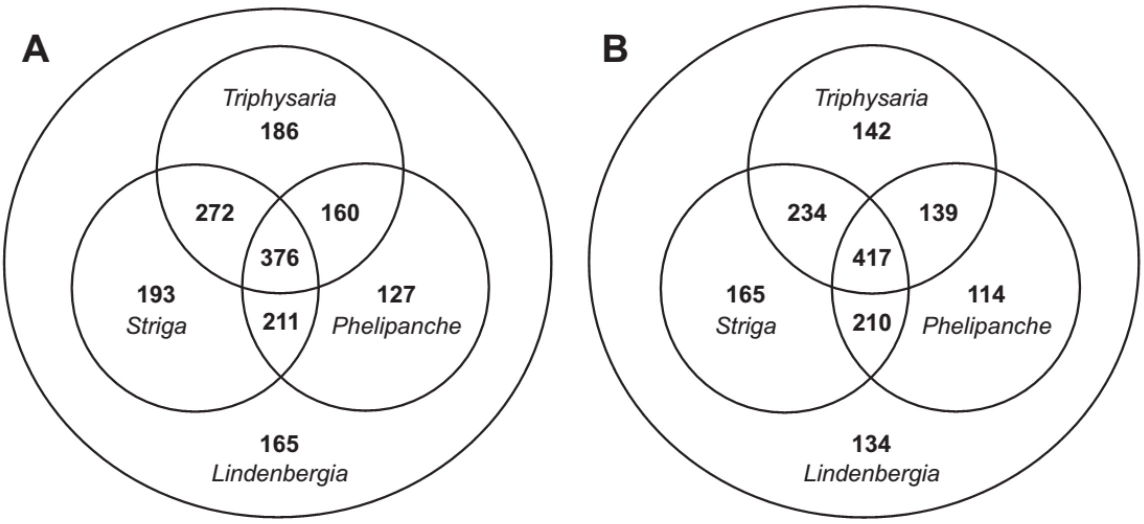


Fig. 2 Distribution of SCGs among the four Orobanchaceae focal species recovered using **a** gapped blast or **b** ungapped blast (see text for details).

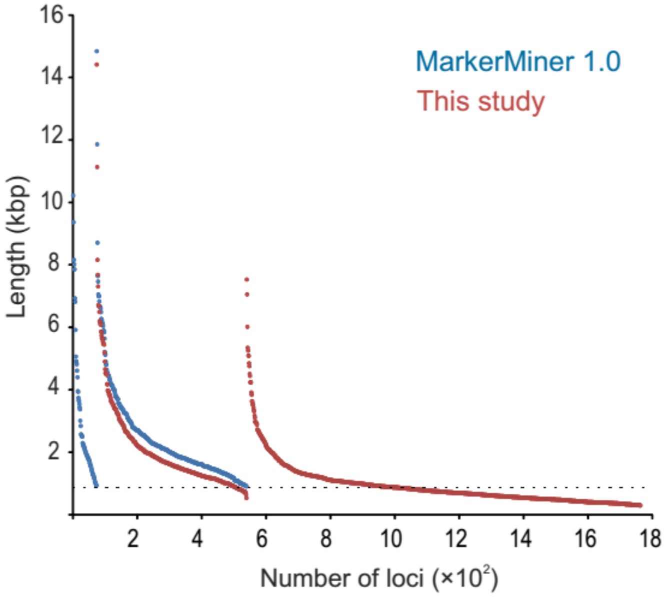


Fig. 3 Length distribution of SCG loci of *Lindenbergia philippensis* identified by MarkerMiner (blue) and by our pipeline (red).

Appendix S1

Minor modifications may be necessary depending on the features of the used datasets. Unless otherwise noted, for the used python scripts the name(s) of the input and the output file(s) have to be changed within the script.
At several steps, BLAST searches are conducted; BLAST software is available from http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastNews.

Descriptions of the pipeline steps adhere to the following format:

1.1 Description of task

```
Command(s)
```

> *Note*: optional; further explanations and/or comments on script and/or program usage and/or on input and output files.
>
> Ex.: example(s)
>
>> *Note*: optional; explanations and/or comments on the presented example

1. Identifying single (low) copy genes (SCGs) in species that are closely related to the group of interest and for which whole genome data are available (henceforth called reference species)

1.1. Generate a database of SCGs from one or more well-annotated genomes (henceforth called master species, e.g., *Arabidopsis thaliana*):

```
formatdb -i input_file -p T
```

Ex.: formatdb -i arabidopsis_single_copy_genes.fasta -p T

Note: "arabidopsis_single_copy_genes.fasta" was obtained from http://cgpdb.ucdavis.edu/COS_Arabidopsis/

1.2. Run BLASTP search with the reference species as query:

```
blastall -p blastp -d subject –i query -e 1e-10 -o output_file
-F "m S" -v 24 -b 24 -g T -m 8
```

Note: It is desirable to obtain maximally long alignments of subject and query sequences. Therefore, to prevent breaking long alignments into several smaller alignments due to intervening low-complexity regions soft masking (-F "m S"), which masks low-complexity regions in the word seeding, but not the extension phase, is used. Additionally, gaps are allowed (-g T) to accommodate length differences of genes in more or less distantly related focal and reference species.

Ex.: blastall -p blastp -d arabidopsis_single_copy_genes.fasta –i Mguttatus_256_v2.0.protein.fasta -e 1e-10 -o Mguttatus.arabidopsis -F "m S" -v 24 -b 24 -g T -m 8

Note: Data for *Erythranthe guttata* (synonym: *Mimulus g.*) were obtained from Phytozome 10.3 (http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PhytozomeV10).

1.3. Filter redundant entries of query sequences (i.e., from the reference species) in the output that are due to different translations (i.e., alternative splicing isoforms), retaining only the longest isoform (“primary transcript”) using a custom python script:

```
1.3_removeX.p.py
```

Ex: "Mguttatus.arabidopsis", obtained in the previous step, as input file, creating "Mguttatus.arabidopsis.1-p" as output file

1.4. Remove query sequences (i.e, from the reference species) with more than one hit with sufficiently high identity scores using a custom python script:

```
1.4_id25_0207sort.single.p.py
```

Note 1: This is achieved in two steps: in the first step, genes of the reference species with identity to genes from the master species (here *Arabidopsis*) of maximally 25% are removed; in the second step, genes from the reference species with more than one hit to the master species are removed.

Note 2: This script requires two input files: one containing the cleaned query sequences generated in step 1.3, the second the protein sequences of the reference species. This script creates two output files, one containing a list of proteins from the reference species that have only a single significant hit to a gene from the master species, the second containing the IDs of these sequences.

Ex.: "Mguttatus.arabidopsis.1-p ", obtained in the previous step, and "Mguttatus_256_v2.0.protein.fasta" (downloaded previously from Phytozome) as input files, creating "Mguttatus.arabidopsis.single.gene.list.fa" and "Mguttatus.arabidopsis.single.gene.list " as output files.

1.5. Remove putatively duplicated genes from the reference species using BLAST and a custom python script.

```
formatdb -i list_of_putative_SCGs_created_in_previous_step -p
T
```

```
blastall -p blastp -d
list_of_putative_SCGs_created_in_previous_step -i query -e 1e-
10 -o output_file -F "m S" -v 24 -b 24 -g T –m 8
```

```
1.5_id25_0207sort.single.p-2.py
```

Note 1: In a first step, reference sequences are blasted against each other. In a second step, query sequences with more than one hit with sufficiently high identity are removed. The thus created list contains the single copy genes from the reference species.

Note 2: This script requires three input files: the first one contains the results from the previous BLAST, the second contains the filtered list of genes from the reference species obtained in step 1.3, and the third contains the protein sequences of the reference species. The script creates two output files, one containing a list of SCGs from the reference species, the second containing the IDs of these sequences. The script uses the same two step procedure as described for step 1.4.

Ex.: formatdb -i Mguttatus.arabidopsis.single.gene.list.fa -p T

blastall -p blastp -d Mguttatus.arabidopsis.single.gene.list.fa -i Mguttatus.arabidopsis.single.gene.list.fa -e 1e-10 -o Mguttatus.vs.Mguttatus -F "m S" -v 24 -b 24 -g T –m 8

"Mguttatus.vs.Mguttatus" (i.e., the list of SCGs from the reference species), "Mguttatus.arabidopsis.1-p" (i.e., the list of genes from the reference species after removal of alternative splice isoforms), and "Mguttatus_256_v2.0.protein.fasta" (i.e., the protein sequences of the reference species downloaded from Phytozome) as input files, creating "Mguttatus.arabidopsis.single.gene.list.-2.fa" and "Mguttatus.arabidopsis.single.gene.list.-2" as output files.

OPTIONAL: If additional reference species are to be included, a joint list of single copy genes can be created, where homologous genes from different reference species are present only once. The following steps are needed:

1.6. Obtain a list of single copy genes from additional reference species following steps 1.1 to 1.5.

Note: Step 1.3 is not necessary, if there are no alternative splicing isoforms indicated.

Ex.: Applying steps 1.1, 1.2, 1.4 and 1.5 to *Solanum lycopersicum* (using "ITAG2.4_proteins.fasta" downloaded from the Sol Genomics Network: ftp://ftp.solgenomics.net/genomes/Solanum_lycopersicum/annotation/ITAG2.4_release/) results in "ITAG.arabidopsis.single.gene.-2.fa", a list of single copy genes in *S. lycopersicum*, and "ITAG.arabidopsis.single.gene.list.-2", containing the IDs of these sequences.

1.7. Identify unique single copy genes (i.e., those not already present in the primary reference species) using BLAST:

```
formatdb -i single_copy_genes_primary_reference_species -p T
```

```
blastall -p blastp -d
single_copy_genes_primary_reference_species –i
secondary_reference_species -e 1e-10 -o output_file -F "m S" -
v 24 -b 24 -g T -m 8
```

Note: The list of single copy genes from the primary reference species (here: *Erythranthe guttata*) has been obtained in step 1.5.

Ex.: formatdb -i Mguttatus.arabidopsis.single.gene.list.-2.fa -p T

blastall -p blastp -d Mguttatus.arabidopsis.single.gene.list.-2.fa –i ITAG.arabidopsis.single.gene.list.-2.fa -e 1e-10 -o ITAG.-2.vs.Mguttatus.-2 -F "m S" -v 24 -b 24 -g T -m 8

1.8. Remove SCGs from the secondary reference with high similarity to SCGs from the primary reference using a custom python script

```
1.8_removeS.py
```

Note: This script requires two input files: one containing the list of sequences from the second reference having significant hits with sequences from the primary reference (i.e., the blast output file generated in step 1.7), and the other containing the IDs of single copy sequences from the second reference species generated in step 1.6. The script produces as output file a list of IDs of genes from the second reference species that have no hit to genes from the primary reference species.

Ex.: Using "ITAG.-2.vs.Mguttatus.-2", created in step 1.7, and "ITAG.arabidopsis.single.gene.list.-2", generated in step 1.6, to generate the output file "ITAG_nohit".

1.9. Extract protein sequences of the secondary reference species using a python script.

```
1.9_extract proteins.py
```

Note: This script requires two input files: one containing the list of IDs of genes from the second reference species that have no hit to genes from the primary reference species, generated in step 1.8, the second containing the protein sequences of the secondary reference species (already downloaded from Phytozome). This script produces a single output file, containing title and protein sequences of the unique SCGs from the secondary reference species.

Ex.: "ITAG_nohit" and "ITAG2.4_proteins.fasta" as input file, creating "ITAG_sorted" as output file.

1.10. Merge lists of single copy genes from the primary and the secondary references species.

Ex.: "Mguttatus.arabidopsis.single.gene.-2.fa" and "ITAG_sorted" were merged to "Mguttatus_Solycdual.sorted.fa".

1.11. Merge lists of coding sequences (CDS) from the primary and the secondary references species.

Ex.: "Mguttatus_256_v2.0.cds.fasta" (downloaded from Phytozome) and "ITAG2.4_cds.fasta" (downloaded from the Sol Genomics Network) were merged to create "Mguttatus_Solycdual.sorted.cds".

2. Identifying single copy genes in focal species.

Note: We used as focal species four species of Orobanchaceae (from the family sequence baits are developed for), whose data were obtained from the Parasitic Plant Genome Project (PPGP) at http://ppgp.huck.psu.edu/download.php: *Lindenbergia philippensis* ("LiPhGnB1.fasta"), *Triphysaria versicolor* ("TrVeBC2.fasta"), *Striga hermonthica* ("StHeBC2.fasta"), *Phelipanche (Orobanche) aegyptiaca* ("OrAeBC4.fasta").

2.1. Run BLASTX search of unigenes from each of the query taxa (here: the four Orobanchaceae) separately against the reference.

```
formatdb -i reference -p T
```

```
blastall -p blastx -d reference –i unigenes -e 1e-10 -o
output_file -F "m S" -v 24 -b 24 -g F -m 8 OR blastall -p
blastx -d reference –i unigenes -e 1e-10 -o output_file -F "m
S" -v 24 -b 24 -g T -m 8
```

Note 1: The reference is the list of SCGs from the primary and, if applicable, secondary reference species generated in steps 1.1 to 1.5 or to 1.10, respectively.

Note 2: Whether to use ungapped (-g F) or gapped BLAST (-g T) will depend on the study design. As outlined under step 1.2, gapped BLAST will allow for longer alignments and thus result in more data, while ungapped BLAST may be preferable if more conserved baits and fewer data are needed.

Ex.: formatdb -i Mguttatus_Solycdual.sorted.fa -p T

blastall -p blastx -d Mguttatus_Solycdual.sorted.fa -i StHeBC2.fasta -e 1e-10 -o StHeBC2_ungap.out -F "m S" -v 24 -b 24 -g F ***OR*** blastall -p blastx -d Mguttatus_Solycdual.sorted.fa -i StHeBC2.fasta -e 1e-10 -o StHeBC2_gap.out -F "m S" -v 24 -b 24 -g T

2.2. Filter out data with insufficient quality using tcl_blast_parser_123.

```
tcl_blast_parser_123_V047.tcl input_file output_file
expect_cutoff identity_cutoff overlap_cutoff matrix_option
```

Note 1: tcl_blast_parser_123_V047.tcl is available from http://code.google.com/p/atgc-tools/downloads/detail?name=tcl_blast_parser_123_V047.tcl

Note 2: This blast parser produces several output files (for details see http://www.atgc.org/BlastParser/Blast_Parser_017.html), of which only the *matrix* file (contains information on primary hits only organized in five columns with query ID, subject ID, identity (normalized between 0 and 1), normalized expectation, and alignment length, respectively) and the *all_hits* file (contains information on all hits organized in 14 columns, including those pertaining to start and end points of the alignments of query and subject) are needed for the purpose of this pipeline.

Note 3: The argument *expect_cutoff* is given as –log(e-value). The last argument, *matrix_option*, indicates whether BLAST file processing should end at the matrix level (argument “MATRIX”) or at the graph level (argument “GRAPH”), i.e., after sequence clustering has been done. The latter can be very time-consuming and is not necessary for the purpose of this pipeline.

Ex: tclsh tcl_blast_parser_123_V047.tcl StHeBC2_ungap.out StHeBC2_ungap.out 20 40 100 MATRIX

Note: The matrix file "StHeBC2_ungap.out.matrix.identity" contains primary hits if they satisfy each of the following conditions: identity at least 40, expectation better than 1e-20 (the cut-off is given as –log(e-value), i.e., 20), and an alignment length of at least 100 positions (at the amino acid level). The second output file of interest is "StHeBC2_ungap.out.all_hits”.

2.3. Combine the *all_hits* files of each query taxon (here: the four Orobanchaceae) into a single file using a custom python script.

```
2.3_merge_all_hits.py
```

Note: This script requires a folder named “raw”, which contains the *all_hits* files, as input. The script first generates a folder named “clean”, containing the *all_hits* file with unigenes without any hit to the reference being removed, and then a folder named “combine”, where the cleaned *all_hits* files have been combined to a single file “4_all_hits_nogap”.

Ex.: Using *all_hits* files of each query taxon ("LiPhGnB1_ungap.out.all_hits", "TrVeBC2_ungap.out.all_hits", "StHeBC2_ungap.out.all_hits", "OrAeBC4_ungap.out.all_hits") in folder "raw" as input, generating folders "clean" (containing filtered *all_hits* files) and "combine" (containing a combined file of the filtered *all_hits* files) as output.

2.4. Remove putative contamination from host species.

Note: This step is only necessary for parasitic species (here: *Triphysaria versicolor*, *Striga hermonthica*, *Phelipanche (Orobanche) aegyptiaca*), but not for non-parasitic ones (here: *Lindenbergia philippensis*). The combined builds from the parasitic species available from PPGP may include sequences from the host species, which have to be removed. If the precise host species the parasite has been grown on is not known, hosts reported in the literature may be used for filtering.

2.4.1. Parasite unigenes are blasted against host data.

```
formatdb -i host_species -p T
blastall -p blastn -d host_species -i parasite_species -e 1e-
100 -o output_file -v 24 -b 24 -g T
```

Ex.: formatdb -i Sbicolor_255_v2.1.cds.fa -p T

blastall -p blastn -d Sbicolor_255_v2.1.cds.fa -i StHeBC2.fasta -e 1e-100 -o StHeBC2.sort_Sb.out -v 24 -b 24 -g T

Note 1: Using gapped BLAST (-g T) makes this approach more conservative by allowing more deviation from reference host data.

Note 2: *Triphysaria versicolor* ("TrVeBC2.fasta") was blasted aginst *Medicago truncatula* ("Mtruncatula_285_Mt4.0v1.cds.fa", downloaded from Phytozome); *Striga hermonthica* ("StHeBC2.fasta") was blasted against *Sorghum bicolor* ("Sbicolor_255_v2.1.cds.fa", downloaded from Phytozome); *Phelipanche (Orobanche) aegyptiaca* ("OrAeBC4.fasta") was blasted against *Arabidopsis thaliana* ("Athaliana_167_TAIR10.cds.fa", downloaded from Phytozome) and *Nicotiana tabacum* ("Ntab-TN90_AYMY-SS_TN90.cds.annot.fna", downloaded from ftp://ftp.solgenomics.net/genomes/Nicotiana_tabacum/annotation/).

2.4.2. Construct a joint list of putative host sequences using a custom python script:

`2.4.2_build_hostlist.py`

Note 1: Parasite sequences with identity to host sequences of at least 95% are regarded as host contaminations.

Note 2: This script requires a single input file containing the blast results of parasite unigenes against host sequences generated in step 2.4.1. This script generates a single output file, containing a list of IDs of parasite sequences that are assumed to be host sequences.

Ex.: Using "StHeBC2.sort_Sb.out" as input, generating "StHeBC2.sort_Sb.out.sort" as output.

2.4.3. Merge host sequences into a single file using a python script:

`2.4.3_FileMerge.py`

Note: This script is available from http://stackoverflow.com/questions/17749058/combine-multiple-text-files-into-one-text-file-using-python and combines query IDs from all three parasitic Orobanchaceae that are assumed to be host sequences generated in 2.4.2. into a single file, "all_host".

Ex.: Using "StHeBC2.sort_Sb.out.sort", "TrVeBC2.sort_Mt.out.sort", "OrAeBC4.sort_At.out.sort" and "OrAeBC4.sort_Nt.out.sort" as input, generating "all_host" as output.

2.4.4. Remove host sequences using a custom python script:

`2.4.4_remove_host.py`

Note: This script requires two input files: one containing the list of all host sequences generated in step 2.4.3, the second containing the *matrix* file generated in step 2.2. This script generates a single output file, i.e., a *matrix* file where putative host sequences have been removed.

Ex.: Using “all_host " and "StHeBC2_ungap.out.matrix.identity " as input, generating " st_ungap_host_free" as output.

2.5. Identify and remove putative paralogues.

Note: Different unigenes from the same query taxon (here: any of the four Orobanchaceae) that have been blasted against the same reference single copy gene with a user-defined overlap cut-off are considered as putative paralogues.

2.5.1. Add position information to *matrix* files using a custom python script:

`2.5.1_add_coordinate.py`

Note 1: The *matrix* files generated in steps 2.2. or 2.4.4., respectively, do not contain information on the alignment positions, which is, however, available in the *all_hits* files.

Note 2: This script requires two input files: one is the *matrix* file (for non-parasites the one(s) obtained in step 2.2., for parasites those obtained in step 2.4.4.), the second one is the combined *all_hits* file obtained in step 2.3. This script produces one output file, corresponding to a *matrix* file with four additional columns pertaining to alignment start and end points of query and subject sequences.

Ex.: Using "st_ungap_host_free", generated in step 2.4.4. and "4_all_hits_nogap", generated in step 2.3. as input files, to generate the output file "st_ungap_host_free.add".

2.5.2. Identify overlapping query sequences that blast against the same reference gene (thus indicating putative paralogy) using a custom python script.

Note: The extent of permitted overlap determines how strictly the paralogy filter is applied. We use here a permitted overlap of maximally 3 bps, thus removing more unigenes especially from those Orobanchaceae with comparatively poor data (i.e., many short unigenes as, at least partly, an artificial result of assembly), such as *Triphysaria versicolor*.

```
2.5.2_note_paralogs.py
```

Note: This script uses a single input file corresponding to the extended *matrix* file generated in step 2.5.1. In a first step, entries of this input file (containing data from one focal species only) are sorted in ascending order using firstly the gene-ID of the reference species (i.e., the subject ID, in the fourth column) and secondly the alignment start point of the subject sequence (i.e., from the reference species; in the fifth column), creating a new output file. In a second step, for each gene of the reference species it is checked whether the alignment end point in row *n* is larger by at least 3 than the alignment start point in row *n*+1: if yes, then the query sequences (i.e., the Orobanchaceae unigenes) in rows *n* and *n*+1 are considered putative paralogues, which is indicated with an entry in the last column of the thus created second output file.

Ex.: Using "st_ungap_host_free.add" as input file to generate "st_ungap_host_free.add.sorted" (i.e., the sorted file) and "st_ungap_host_free.add.sorted.paralogInfo" (the sorted file with indication of putative paralogy in the ninth column) as output files.

2.5.3. Remove putative paralogues using a custom python script:

```
2.5.3_remove.py
```

Note: This script uses as input file the sorted list with indication of putative paralogy generated in step 2.5.2 and creates a single output file, where putative paralogues have been removed. If any of the query sequences (i.e., unigenes) sharing the same reference gene (i.e., subject) is indicated as putative paralogue, the entire set of unigenes aligned to the same reference gene is removed from the focal species.

Ex.: Using "st_ungap_host_free.add.sorted.paralogInfo" as input to create "st_ungap_host_free.add.sorted.paralogInfo.out" as output file.

2.5.4. Combine the paralogue-free data (i.e, the output files generated in the 2.5.3) into a single file using a python script.

```
2.4.3_FileMerge.py
```

Note: This is the same python script used in step 2.4.3.

Ex.: "st_ungap_host_free.add.sorted.paralogInfo.out", "tr_ungap_host_free.add.sorted.paralogInfo.out", "or_ungap_host_free.add.sorted.paralogInfo.out" and "li_ungap_host_free.add.sorted.paralogInfo.out" were combined to generate "4_nogap_hostfree_3sorted".

3. Extract sequences from both reference and focal species

3.1. Extract sequences of the non-paralogues genes from each of the focal species using a custom python script.

```
3.1_cut-seq.py
```

Note: This script requires two input files: one containing the paralogy-free data obtained in step 2.5.4, the second containing the protein sequences from the focal species downloaded previously. This script generates a single output file containing only the aligned part of the sequences from the focal species.

Ex.: Using "4_nogap_hostfree_3sorted" and "StHeBC2.fasta" to generate "StHeBC2_nogap_edit.codons.query_all.seq".

3.2. Combine, per gene, sequences of the reference species and of all focal species into a single file (i.e., one file per locus) using a custom python script:

```
3.2_query-seq.py
```

Note: This script requires at least three input files: One containing the paralogy-free data obtained in step 2.5.4 (containing start and end points of reference and focal species), the second containing the coding sequences of the reference species generated in steps 1.1–1.5 (or 1.1–1.11), the third one containing the sequences of the non-paralogous genes of the focal species obtained in step 3.1 (if more than one focal species is used, the files for all species can be loaded simultaneously, here resulting in six instead of three input files). This script creates an output folder named "group1" containing one fasta file per gene. Each fasta file contains the newly extracted gene sequence from the reference species (start and end point are the minimum start and maximum end point over all alignments to sequences from the focal species, i.e., this sequence from the reference species may contain stretches without any aligned positions from the focal species) and all aligned sequences from the focal species.

Ex.: Using "4_nogap_hostfree_3sorted " (i.e., containing the paralogy-free data obtained in step 2.5.4), "Mguttatus_Solycdual.sorted.cds" (i.e., coding sequence data of the reference species), as well as "StHeBC2_nogap_edit_codons.query_all.seq", "OrAeBC4_nogap_edit _codons.query_all.seq", "TrVeBC2_nogap_edit _codons.query_all.seq", and "LiPhGnB1_nogap_edit _codons.query_all.seq" (i.e., the four files containing the sequences of the non-paralogous genes of each of the focal species obtained in step 3.1) to create numerous fasta files (one per gene) in the folder "group1".

3.3. Bring sequences to the same orientation relative to the reference species using a custom python script:

```
3.3_reverse-negative-seqs.py
```

Note: As some (or all) of the sequences from the focal species may be aligned to the reference species as reverse complement, re-orientation is necessary prior to the alignment to be conducted in step 4. This script goes through the fasta files created in step 3.2 and produces reverse complements for those sequences of the focal species, whose BLAST alignment endpoint on the reference species is smaller than the start point. The thus altered fasta files are written into a new folder "group2".

Ex.: Using "group1", generated in step 3.2, as input folder, the output folder "group2" containing one fasta file per gene is generated.

OPTIONAL: If data quality from some of the focal species is non-satisfactory (e.g., too many short unigenes), data from genes that (apart from the always present reference species) only contain such focal species may be considered insufficiently reliable and may be removed. Here, the single highest quality focal species is *Lindenbergia philippensis*, which is used to retain only those genes where (in addition to the reference species) data from at least *Lindenbergia* are present.

3.4. Retain only genes where data from focal species with sufficient data quality (here *Lindenbergia philippensis*) are present using a custom python script:

```
3.4_sort-by-LiPhGnB.py
```

Note: This script uses as input the folder created in step 3.3 and generates a new folder "group3" containing only fasta files for those genes that contain data from suitable focal species.

Ex.: Using "group2" as input folder, creating "group3" as output folder.

4. Sequence alignment

Note: When identifying single copy genes from the focal species (as described in section 2), in step 2.1 either ungapped or gapped BLAST can be used. In case of ungapped BLAST, sequences from reference and focal species can be aligned unambiguously, as there are no length differences between aligned segments; consequently, also sequences from focal species can be aligned unambiguously, although sequences from different focal species may only partly overlap or even not overlap at all, as they align to different sections of the sequence from the reference species. For gapped BLAST, putatively present length differences between aligned sequences of reference and focal species may cause ambiguities in the alignment not only between reference and focal species, but especially among focal species. Therefore, a dedicated alignment program is used to re-align all sequences. Accordingly, different steps are required, which will be described in turn ("u" indicating steps necessary in case of ungapped BLAST, "g" indicating steps necessary in case of gapped BLAST).

4.u1. Align sequences of the focal species and the reference species using information on start and end points of the BLAST alignment using a custom python script:

```
4.u1_filling-in1-title.py
```

Note 1: This script does two things: First, it aligns sequences from focal and reference species based on the BLAST alignment by adding gaps at the beginning and/or the end of the shorter sequence; second, white spaces in the sequence titles are replaced with asterisks (necessary for later steps), i.e., "Subject_ID Subject_start_site-Subject_end_site length_of_Subject" becomes "Subject_ID*Subject_start_site-Subject_end_site*length_of_Subject" (e.g., ">LiPhGnB1_97231 546-1703 forward [Migut.A00009.1*232-1389] length=1158" becomes ">LiPhGnB1_97231*546-1703*forward*[Migut.A00009.1*232-1389]*length=1158").

Note 2: The script requires an input folder, created in step 3.4, and creates an output folder, containing fasta files of aligned sequences.

Ex.: Using "group3" as input folder, creating "group4" as output folder.

4.g1. Align sequences of the focal species and the reference species using MAFFT called by a custom python script:

```
4.g1_python_call_mafft.py
```

Note 1: The script assumes that the MAFFT executable is either in the folder the script is run from or that the location of MAFFT is in the PATH variable. For alignment, we have chosen as iterative refinement method E-INS-I (i.e., options --ep 0 --genafpair --maxiterate 1000), which is suitable for data sets containing multiple conserved domains and long gaps.

Note 2: The script requires an input folder, created in step 3.4, and creates an output folder "mafft_out" containing fasta files of aligned sequences.

Ex.: Using "group3" as input folder, creating "mafft_out" as output folder.

4.g2. Change the MAFFT output from multi-line to single-line fasta using a custom python script:

```
4.g2_change-multe-fasta.py
```

Note: This script takes an input folder containing the fasta files with the MAFFT-aligned sequences created in the previous step 4.g1 and creates an output folder containing the modified fasta files.

Ex.: Using "mafft_out" as input folder creating "group4.0" as output folder.

4.g3. Changing white spaces in the sequence titles to asterisks (necessary for later steps) using a custom python script:

```
4.g3_Gap-filling-in1-title.py
```

Note 1: The script requires an input folder, created in step 4.g2, and creates an output folder, containing fasta files of aligned sequences with modified sequence names.

Ex.: Using "group4.0 " as input folder, creating "group4" as output folder.

5. Extract exon sequences and remove those of insufficient length

Note: Exon sequences were extracted based on information on exon limits available in the gff3-files of the reference species, here "Mguttatus_256_v2.0.gene.gff3" and "Slycopersicum_225_iTAGv2.3.gene.gff3" (both downloaded from Phytozome10.3: http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PhytozomeV10)

5.1. Extend sequences to start with the first position using a custom python script:

```
5.1_filling-in2-start-coordinate-sites.py
```

Note 1: Sequences from the aligned reference species may not start with the first position of the entire coding region (cds). To simplify subsequent steps, the sequences from the reference and the focal species are extended at the beginning by inserting plus signs making alignment position 1 corresponding to cds position 1;

e.g., >Migut.A00009.1 starts at cds position 232, accordingly 231 "+" have to be inserted at the beginning of this sequence, from the reference species, and those from the focal species aligned to it.

Note 2: This script takes the folder containing the aligned sequences created in steps 4.u1 or 4.g3, respectively, as input and creates an output-folder containing the extended aligned sequences.

Ex.: Using "group4" as input folder creating "filling-in" as output folder containing fasta files with extended aligned sequences.

5.2. Transpose the aligned sequences using a custom python script:

`5.2_transfer-x2y.py`

Note 1: To make subsequent steps easier, the sequence alignments are transposed. The resulting file has as many columns as aligned sequences, its first row contains the sequence names and rows 2 to n+1 contain the first to n^{th} alignment position.

Note 2: This script takes the folder containing the extended aligned sequences created in step 5.1 as input and creates an output-folder containing the transposed alignments.

Ex.: Using "filling-in" as input folder creating "group5" as output folder containing the transposed alignments.

5.3. Clean the files with transposed sequences using a custom python script:

`5.3_clean.py`

Note: This script takes the folder containing the transposed alignments created in step 5.2 as input and creates an output-folder containing the transposed alignments, where empty lines, if present in the input, have been removed.

Ex.: Using "group5" as input folder creating "group6" as output folder containing the cleaned transposed alignments.

5.4. Compile start and end points of all exons per gene in a single file using a custom python script:

`5.4_gene-map.py`

Note: This script uses gff3-files from the reference species as input files and creates a single output file containing gene name, gene orientation, and start and end points of the gene’s exons (one line per gene).

Ex.: Using "Mguttatus_256_v2.0.gene.gff3"and "Slycopersicum_225_iTAGv2.3.gene.gff3" downloaded previously as input files, creating "map" as output file.

5.5. Add flags for split site positions in the transposed alignments using a custom python script:

```
5.5_add-split-signal.py
```

Note: This script requires an input folder, containing the cleaned transposed alignments generated in step 5.3, and an input file containing the list of start and end points of exons created in step 5.4. The script creates an output folder containing the transposed alignments, where additionally the end-points of each exon are indicated by an additional line containing an "S" for each sequence.

Ex.: Using "group6" as input folder and "map" as input file, creating "group7" as output folder.

5.6. Extract exons into separate files using a custom python script:

```
5.6_split.py
```

Note: This script takes the folder containing the transposed alignments with split site positions indicated created in step 5.5 as input and creates an output-folder containing separate transposed alignment files from each exon.

Ex.: Using "group7" as input folder as input file, creating "group8" as output folder. File "Migut.B00613.1", pertaining to a gene with 3 exons, will be separated into "Migut. B00613.1.1", "Migut. B00613.1.2" and "Migut. B00613.1.3".

5.7. Remove exons of insufficient length using a custom python script:

```
5.7_sort120.py
```

Note 1: The desired bait length depends on the study design and may vary from 80 to 120 bp; the length threshold below which an exon is removed should be adjusted accordingly. Here, we are interested in longer baits and, hence use 120 bp as cut-off.

Note 2: This script takes the folder containing the exon alignments created in step 5.6 as input and creates an output-folder containing only those exon alignments that exceed a certain threshold.

Ex.: Using "group8" as input folder, creating "group9" as output folder. If lengths of the three exons "Migut.A00009.1.1", "Migut.A00009.1.2" and "Migut.A00009.1.3" are 210 bp, 105 bp and 300 bp, only "Migut.A00009.1.1" and "Migut.A00009.1.3" will be retained.

6. Extract baits

6.1. Re-transpose files containing the exon sequences to fasta files using a custom python script:

`6.1_transfer-y2x.py`

Note 1: Alignments have been transposed in step 5.2 to make extraction of exons easier. From here on, the transposed orientation is no longer needed and is, therefore, reverted.

Note 2: This script takes the folder containing the alignments of sufficiently long exons created in step 5.7 as input and creates an output-folder containing only those exon alignments in standard fasta format.

Ex.: Using "group9" as input folder, creating "group10" as output folder.

6.2. Remove gaps in exons using a custom python script:

`6.2_degap.py`

Note 1: In order to obtain baits that might be targeting different focal species, gaps in the exons have to be removed.

Note 2: This script takes the folder containing the alignments of sufficiently long exons in fasta format created in step 6.1 as input and creates an output-folder containing the alignments of sufficiently long exons without gaps.

Ex.: Using "group10" as input folder, creating "group12" as output folder.

Note: An intermediary results folder, "group11", is also created, which differs from "group12" only that names in each file do not have the suffix "120bp", but may be ignored.

6.3. Extract baits using a custom python script:

`6.3_final-split-baits.py`

Note 1: Length of baits (here 120 bp) and tiling density (here 2×, i.e., a 60 bp overlap) depend on the study design. To avoid redundancy between baits, a threshold is set for the maximally permitted overlap of baits (here 80 bp, i.e., from a fragment of 160 bp still two baits, overlapping by 80 bp, could be extracted), but this threshold may be adjusted if needed.

Note 2: In the first step, this script takes the folder containing the alignments of sufficiently long exons without gaps created in step 6.2 as input and creates an output-folder "group13" containing the baits from one exon of one focal species (e.g., a gene with five exons and each exon with data from three focal species will result in 15 output files). In the second step, all baits of the focal species are combined into a single fasta file (i.e., the folder containing the baits, "group13", is taken as input, resulting in single file as output).

Ex.: Using "group12" as input folder, creating "group13" as output folder; using "group13" as input folder, creating "nogap_baits.combined.fasta" as output file.

6.4. Combine exon sequences into single files per species using a custom python script:

```
6.4_sort-4sp.py
```

Note: This script combines all exons of a single focal species into one file. It takes the folder containing the alignments of sufficiently long exons without gaps created in step 6.2 as input and creates a single fasta file per species containing all exons.

Ex.: Using "group12" as input folder, creating "StHe.fasta","OrAe.fasta", "TrVe.fasta", and "LiPh.fasta" as output file.

7. Data cleansing

7.1. Remove genes with too few baits using a custom python script:

```
7.1_length280.py
```

Note 1: Too short loci that should not be targeted have to be removed. The threshold, below which a locus is considered too short, will depend on the study design. Here we use the number of baits per species as criterion: although the number of baits will also depend on exon structure (i.e., a gene with many short exons will yield fewer baits than a gene with a single long exon, even if the total length of the cds is the same), the number of baits will correlate with gene length. Here, we use a cut-off value of four baits (corresponding to at least 280 bp per gene, i.e., the minimum length if the baits are from a single exon and the last bait has an overlap of 80 bp with the preceding bait), but this cut-off value may be adjusted depending on the study design. As data quality differs among our focal species, we only considered a single focal species, *Lindenbergia philippensis*, which has the best quality data (i.e., the longest unigenes). In consequence, a gene is removed, if there are fewer than four baits from *L. philippensis*, irrespective of the number of baits available for this gene from any of the other focal species.

Note 2: This script takes the fasta file containing the baits per species created in step 6.3 as input and creates an output file containing only baits for those genes that have at least four baits for *L. philippensis*.

Ex.: Using "nogap_baits.combined.fasta" as input file, creating "no_gap_sorted.result.fa" as output file.

7.2. Remove redundant baits (i.e., having a similarity above a certain threshold) using CD-HIT-EST.

Note: We used online CD-HIT-EST (available at: http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit-est). As baits from different genes, but with similar sequences, may cause cross-hybridization during enrichment, only one of those should be retained. The precise value of the cut-off depends on study design; here, we used ≥90% (minIdentity=90; other settings were left at their default).

Ex.: "no_gap_sorted.result.fa" obtained in previous step as input file, generating "1456519283nogap.fas.1" as output file containing baits that share less than 90% identities.

7.3. Identify baits with high similarity to plastid sequences using blast.

```
formatdb -i baits -p T
```

```
blastall -p blastn -d baits –i plastid_genomes -e 1e-10 -o
output  -v 24 -b 24 -g T -m 8
```

Note: It is very important to remove baits that are similar to plastid sequences, because these are present in high copy numbers and will be highly enriched during the enrichment phase, reducing the yield of sequences from single copy nuclear genes. The precise similarity threshold used will depend on the study design. Here we blast baits against 12 plastid genomes (listed below).

Ex.: formatdb -i 1456519283nogap.fas.1-p T

blastall -p blastn -d 1456519283nogap.fas.1–i cpgenome.fasta -e 1e-10 -o 1456519283nogap.fas.1.out -v 24 -b 24 -g T -m 8

Note: We used 12 plastid genomes: *Boulardia latisquama*, *Cistanche deserticola*, *Conopholis americana*, *Epifagus virginiana*, *Lindenbergia philippensis*, *Orobanche crenata*, *Orobanche gracilis*, *Phelipanche purpurea*, *Phelipanche ramosa*, *Schwalbea americana*, *Striga hermonthica*, *Triphysaria versicolor.*

7.4. Remove putative plastid sequence using a custom python script:

```
7.4_remove_plastid.py
```

Note: This script uses two input files, one containing the sorted baits that share less than 90% identities generated in step 7.2 and another one containing the blast output file of baits against plastid genomes generated in step 7.3.This script generates two output files, one containing the list of genes, which have >=90% identities with any one of the 12 plastid genomes, another containing baits where sequences of putative plastid origin have been removed.

Ex.: Using "1456519283nogap.fas.1" and "1456519283nogap.fas.1.out" as input to create "1456519283nogap.fas.list" and "1456519283nogap.fas.1.plastid_free" as output files.

Conclusion

Analysis of a few hand-selected markers, especially if joined with other evidence, for instance, from karyology, can be sufficient to place unknown groups within a phylogenetic framework. This is the case for the enigmatic genus *Platypholis*. Both molecular phylogenetic and karyological data show that this genus phylogenetically belongs to the *Orobanche* clade and is sister to or even nested within *Orobanche* s. str, although its phylogenetic placement within *Orobanche* s. str. is less certain. A closer relationship to *O. coerulescens*, as suggested by ITS data, is supported by geographic proximity. The considerably larger genome observed in *Platypholis* may be connected to life-history or changes in breeding system. The uncertainty concerning the precise placement of *Platypholis* may be due to contradict phylogenetic relationships inferred from plastid versus nuclear markers that resulted from ILS, which can be accommodated by using species tree estimation methods in the future. This will, however, require much larger data sets, especially if a possible negative affect of missing data is to be avoided.

To alleviate the issues such as ILS, more markers are needed, a path we followed to improve our understanding of phylogenetic relationships among major clades in Orobanchaceae. It took us plenty of effort to test different PPR and LCN genes, and we confirmed that two PPR genes are well suited to address phylogenetic problems at family or lower levels in Orobanchaceae, while the single LCN markers show less resolving power. Although there are minor conflicts among different markers in some of the major clades, these receive poorly support. Major clades previously identified are mostly confirmed. Against most previous suggestions, the holoparasitic *Orobanche* clade is consistently supported as sister to all parasitic clades. The newly added *Phtheirospermum* species and *Pterygiella* form a separate clade (*Pterygiella* clade).Unclear relationships do, however, remain (e.g., the precise placement of the *Cymbaria-Siphonostegia* clade, *Brandisia*, and relationships among the *Pterygiella* clade and other clades).

More markers are needed though phylogenomic approaches to resolve the unclear relationships, when single or several markers are still insufficient. For the purpose of gaining more orthologous LCN loci to accurately resolve the backbone of Orobanchaceae, we establish a highly efficient bioinformatic pipeline and identified 1307 and 981 SCGs, 38,156 and 21,856 bait sequences, respectively, depending on whether gaps were allowed in initial BLAST searches. Compared to the recently published MarkerMiner 1.0 pipeline, under default setting, ours identified about 1.6 times as many SCGs (of at least 900 bp length). This pipeline is also applicable in other non-model plants, including other parasitic plants, where only transcriptomes or partial genome data of differing quality are available. Empirical demonstration of using our pipeline as a tool in more phylogenomic studies is expected in the future.