**Orchid phylogenetics & herbarium genomics**

**New plastid markers for phylogenetic reconstruction**

**Internship research proposal**

**Julia Faillace Thiesen**

Student number 910131236040

**BSc Biology at Universidade Federal de Santa Catarina**

**Exchange program at Wageningen University**

Internship Thesis proposal

27-3-2014

**5-months Internship project at Wageningen University (no credits)**

**March-July 2013**

**Supervision: Dr. Freek T. Bakker (Biosystematics Group, Wageningen University)**

**External supervisor: Dr. Barbara Gravendeel and Dr. Rutger Vos**

Contents

[1. Introduction 3](#_Toc384797623)

[1.1. State of the Art 4](#_Toc384797624)

[1.2. Plastome Variation & Orchid Phylogenetics 6](#_Toc384797625)

[2. Research Questions 9](#_Toc384797626)

[3. Hypothesis 9](#_Toc384797627)

[4. Materials and Methods 10](#_Toc384797628)

[4.1. Plastome Reconstruction 10](#_Toc384797629)

[4.1.1. Resources Available 10](#_Toc384797630)

[4.1.2. Long Range PCR and Ion Torrent Sequencing 10](#_Toc384797631)

[4.1.3. Gene Annotation 10](#_Toc384797632)

[4.1.4. Submission to GenBank 10](#_Toc384797633)

[4.2. Phylogenetic Analysis 11](#_Toc384797636)

[4.2.1. Whole cpDNA Orchidaceae Alignment 11](#_Toc384797637)

[4.2.2. Phylogeny Inference using intergene plastid regions 11](#_Toc384797638)

[5. Planning 13](#_Toc384797639)

[6. References 14](#_Toc384797640)

# Introduction

Phylogenetic relationships within Orchidaceae are an important subject in current evolutionary biology studies, as the family represents a remarkable range of life history strategies, ﬂoral and vegetative morphology and pollination syndromes (Fay & Chase 2009). Its size and diversity trace a remarkable history of diversification that started 42-36 million years ago (Guo *et al* 2011). Epiphytic habit seems to be one of the stimulating factors for its diversity (Gravendeel *et al*. 2004). Next Generation Sequencing (NGS) technology has delivered fast, inexpensive and accurate genome information (Metzger 2010), and comparative biology has now the opportunity of exploring whole genomes. Whole plastid genome comparisons or concatenated sequences markers have been used for increased phylogenetic resolution at many levels in Angiosperms: ranging from studies up to order level (e. g. within Angiosperms, Chang *et al.* 2006, Shaw 2007, Soltis *et al.* 2011) to species identification and population genomic studies (e. g. Yang *et al.* 2013, Jheng *et al.* 2012, Pan *et al*. 2012, Doorduin *et al.* 2011).

Phylogenetic studies within Cypripedioideae have been done using nuclear and plastid genes, and the topology of the subfamily has become relatively well understood (Guo *et al.* 2012). Contradiction is found in some aspects of the topology. For example, using morphological characters (Pfizter 1903) and genetic data (e.g. plastid *rbc*L gene, Albeit 1994; nuclear nrDNA ITS regions, Cox *et al.* 1997) have placed the genus *Selenipedium* as sister clade to other slipper orchids. On the other hand, studies that used low copy nuclear genes Xdh (Cameron 2006), plastid genes *atpB*, *mat*K and *rbc*L (Freudenstein *et al.* 2003), and a combined dataset using plastid sequences (*mat*K, *rbc*L, *rpo*c1, *rpo*c2, *ycf*1, *ycf*2) and low copy nuclear exons (ACO and Leafy) (Guo *et al* 2012). The position of Cypripedioideae within Orchidaceae is still uncertain based on DNA sequence data comparisons (Stevens 2001 onwards). Figures 1 and 2 show a cladogram of the subfamilies of the Orchidaceae. Figure 1 resumes the topology patterns from Simpson (2010), Cameron *et al*. (2006), Singer *et al.* (2008), Chase *et al.* (2003) and APG tree ( Stevens, P.F. 2001 onwards). The use of historical DNA in genomics have become achievable with NGS methods, and complete genomes of plant and animal collection specimens have been successfully sequenced (Staats *et al.* 2013). Dry-preserved organisms have high levels of DNA degradation, impairing PCR-amplifying reactions that are crucial for Sanger sequencing. NGS approaches, however, do not rely on large DNA fragments but in short molecules of 100-300bp as templates (Staats *et al.* 2013). Reliable sequence data can be obtained from herbarium specimens (Staats *et al* 2011), providing an opportunity to explore historical DNA in a phylogenetic context.

The present study proposes to assembly complete plastome (cpDNA) from two *Paphiopedilum barbatum* (Lindl.) Pfitzer herbarium specimens (as part of a SYNTHESYS project). Aiming to explore rapidly evolving non-coding plastid regions (NCPs) and to detect phylogenetic informative markers, the *P. barbatum* cpDNA will be compared with other available orchid cpDNA. This study will also contribute to orchid phylogetic studies, as it will increase the sampling of whole chloroplast genome within Orchidaceae, and more precisely within Cypripedioideae. Ongoing studies within the subfamily are being conducted by dr. Gravendeel (Naturalis, Leiden; pers. comm.), and other species from the subfamily had their complete cpDNA sequenced and soon are going to be submitted to NCBI GenBank.

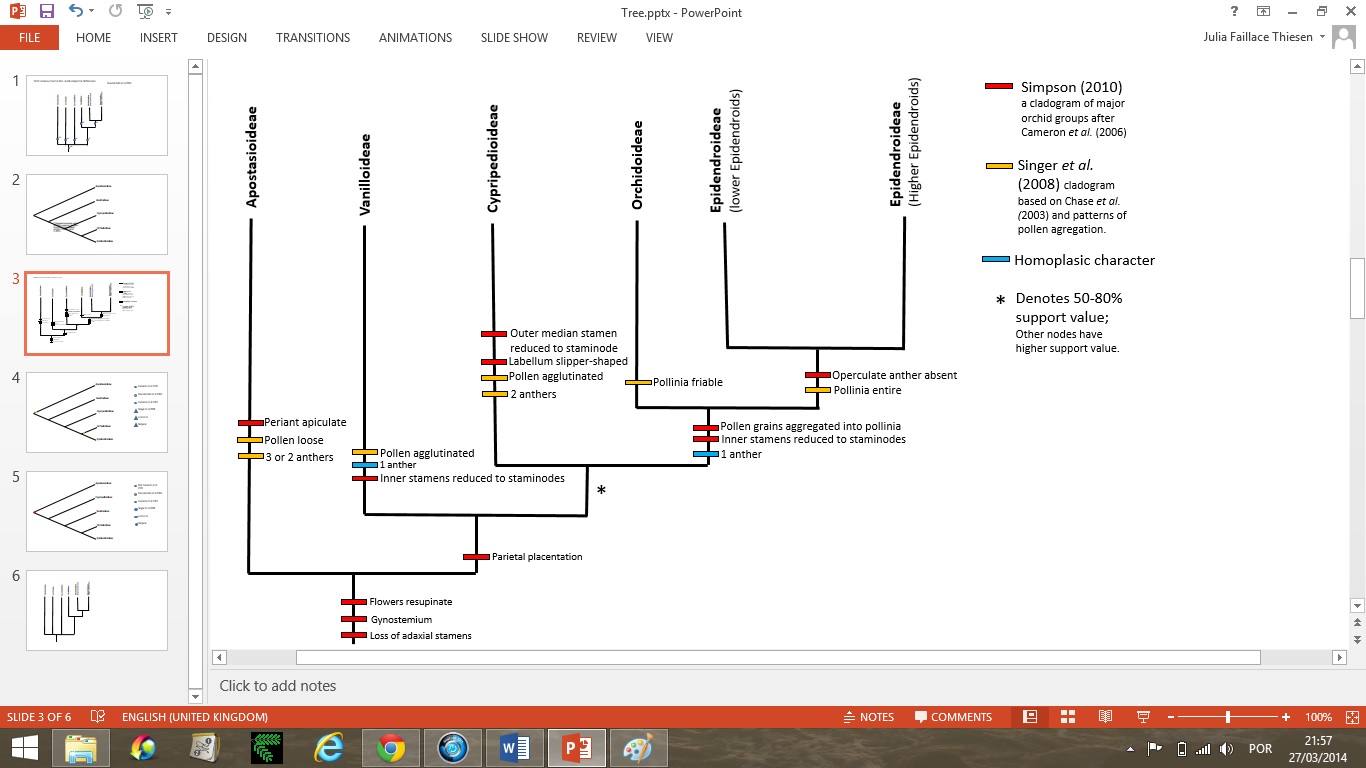
The present study intents to contribute to this study by adding one, possibly two *P. barbatum* full chloroplast genomes, in addition to two already sequenced chloroplast genomes of *Cypripedium calceolus* and *Phragmipedium longifolium* (to be provided by the Leiden team), respectively. These four slipper orchid chloroplast genomes, once assembled, can quite easily be added to an already existing alignment of twelve orchid species from subfamily Epidendroideae. With this alignment expanded with four representatives of the Cypripedioidae, phylogenetic analyses will be carried out to discover promising new plastid markers. I propose to integrate our *P. barbatum* SYNTHESYS samples with dr. Gravendeel’s alignment, and conduct subsequent phylogenetic analysis using RAxML, Maximum Parsimony and MrBayes.

The comparison of 14 complete chloroplast genome using phylogenetic and pair wise distance methods will be useful to detect variable regions among the specimens, and will provide important information in the choice of markers. Nevertheless, 14 terminals may not provide enough information for a phylogenetic reconstruction in Orchidaceae. I propose, therefore, to select the 5 most informative sequences and generate an alignment with the maximum of terminals that are available in GenBank. Phylogenetic reconstruction using MrBayes, Maximum Parsimony and RAxML will be performed. The two phylogenetic trees, one with 14 terminals one with many more, can then be compared.

Recent findings in Biosystematics group (Wageningen University) demonstrate the presence of <1% fungal reads in herbarium samples (Freek pers. comm.) which could be post-mortem contamination of the voucher specimens, or endophytical fungi that used to interact with the specimen. Therefore, the presence of Fungi DNA in *P. barbatum* herbarium specimens will also be explored.

## State of the Art

Orchidaceae is one of the largest families among vascular plants, comprising 800 genera of terrestrial or epiphytic herbs distributed among all continents except Antarctica (Fay & Chase 2009). Beyond their beauty and ornamental value, the range of life history strategies, floral and vegetative morphology and pollination syndromes make them a valuable object for phylogenetic studies. Nevertheless, because of its dimensions, the study of this family in an evolutionary framework is challenging (Chase & Fay 2009): currently there are 26,567 species listed in the World Checklist for the Monocotyledons (2011). Phylogenetic relationships within Orchidaceae are becoming fairly well understood (Singer *et al* 2008), and morphological synapomorphies and

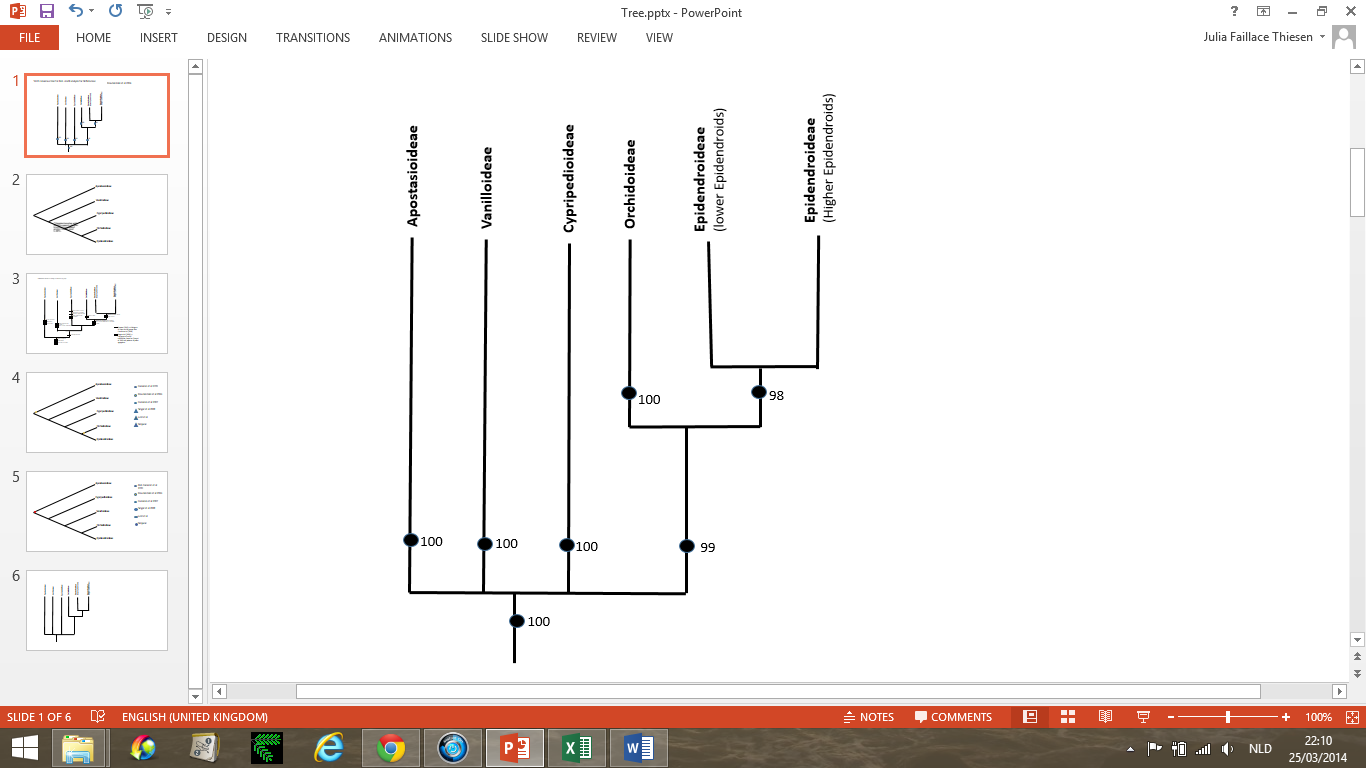


**Figure 1**: Cladogram of major Orchidaceae groups summarizing topology patterns from Simpson (2010),

Cameron *et al*. (2006), Singer *et al.* (2008), Chase *et al.* (2003) and APG tree (Stevens, P.F. 2001 onwards).

The 5 subfamilies from Orchidaceae have been classified based on both molecular data and morphological data; nevertheless there are still uncertainties within the tree (e.g. the placement of Cypripedioideae). One of the hypothesis groups monandrous orchids in one clade (Vanilloideae, Orchidoideae and Epidendroideae). Most recent hypotheses (see Chase *et al.* 2003, Singer *et al.* 2008, Cameron *et al.* 2006) place Cypripedioideae as sister clade to Orchidoideae and Epidendroidea, implying that the loss of one stamen occurred twice (Figure 1).

For many years the phylogeny of the orchids was inferred based almost exclusively on floral features (Chase *et al* 2007). Although morphological data have produced a pertinent hierarchy in Orchidaceae, the occurrence of homoplasy leads to error in many cases (e.g. Chase *et al* 2009, replacement of a set of species in *Gomesa,* previously placed in *Oncidium)*, and may induce errors. In addition, morphology-only-based phylogenies limited the number of characters concatenated and provided less support for clades (Singer et al. 2008). Molecular data have revolutionized phylogenetic studies on Orchidaceae from early 1990s, starting with Chase *et al* (1994) and (Fay & Chase 2009). The first DNA sequence-based studies generated congruent trees (Chase *et al* 1994, Cameron *et al* 1999, Freudenstein *et al* 2004) compared with previous, morphology-based Orchidaceae hypotheses (Dressler 1993). Cameron *et al* (1999) explored the evidences from the plastid gene *rbc*L, using parsimony: 171 taxa across all 5 subfamilies and 1320 characters (485 informative). Comparing with Dressler (1993), the topology generated was mostly congruent, although the elevation of Vanilloid orchids to subfamilial status was an important change to the current topology. Freudenstein *et al* (2004) explored an expanded plastid DNA phylogeny of Orchidaceae using *rbc*L and *mat*K sequences: 173 taxa, 2958 characters (1180 informative). Support values presented in this study are stronger than the previous study (Figure 1). There are now numerous studies that have provided an increasingly detailed phylogenetic framework for orchids (Lay & Chase 2009), and alternative markers are being suggested for its phylogeny (see Yang *et al* 2013, Pan *et al* 2012). Support values for Epidendroideae and Orchidoideae presented in this study are stronger than Cameron *et al.* (1999), but still the topology of the first three diverged subfamilies were not detected, and the tree shows a polytomy for these groups in the strict consensus (Figure 2). A historical overview of phylogenetic studies within Orchidaceae is summarized in Chase *et al* (2007). The authors emphasize important updates about the tree topology and, more importantly, how comparative genomics is transforming and refining taxonomy. *Genera Orchidacearum* (Pridgeon *et al*, 1999; 2001; 2003 up to 2014) is presented as a temporary classification, as new insights are constantly being incorporated.



**Figure 2**: Strict consensus tree for combined rbcL + matK analysis of Orchidaceae (Freudenstein *et al.* 2004).

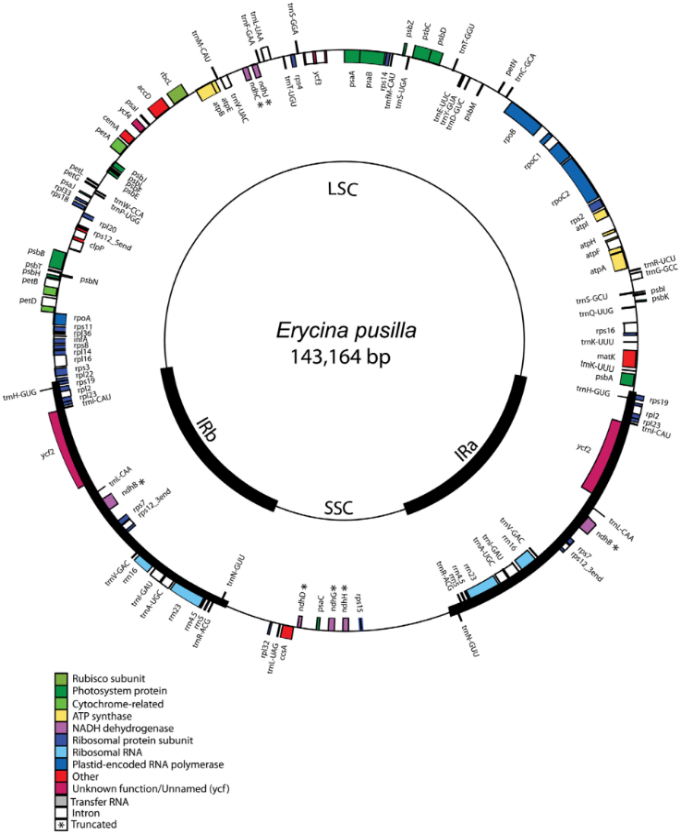
Jackknife support percentages are shown (>50%).

There are now numerous studies that have provided good resolution for lower levels within the subfamilies of Orchidaceae (Lay & Chase 2009), and alternative markers are being suggested for Orchidaceae

phylogeny (see Yang *et al* 2013, Pan *et al* 2012). A historical overview of phylogenetic studies within Orchidaceae is provided in Chase *et al* (2007). The authors emphasize important updates about the tree topology and, more importantly, how comparative genomics is transforming and refining taxonomy. *Genera Orchidacearum* (Pridgeon *et al*, 1999; 2001; 2003 up to 2014) is presented as a temporary classification, as new insights are constantly being incorporated. The progress of sequencing methods have transformed the dimensions in comparative biology, allowing robust alignments of complete genomes and the exploration of better markers for different phylogenetic levels.

## Plastome Variation & Orchid Phylogenetics

Next Generation Sequencing has been used to sequence whole chloroplast genomes in vascular plants, and we can find in GenBank whole chloroplast genomes (cpDNA) from Epidendroideae: *Phalaenopsis aphrodite* (Chang *et al.* 2006), *Phalaenopsis equestris* (Jheng *et al.* 2012), *Erycina pusilla* (Pan *et al* 2012), *Cymbidium* spp. (Yang *et al.* 2013), Oncidium Gower Ramsey (Wu *et al.* 2010). *Rhizanthella*

*gardneri,(*Delannoy *et al.* 2011) plastome sequence, from Orchidoideae, can also be found in GenBank. These authors have explored the plastid genome, finding that besides its conserved overall structure there are phylogeneticaly informative hotspots (e.g. different pattern of *ndh* subunits loss and truncation, indels patterns of *ndh* peseudogenes, IR junctions, SNPs and certain intergenic spacers).

**Figure 3**: Simplified from Pan *et al.* (2012). Quadripartite plastome structure of *Erycina pusilla* (Orchidaceae).

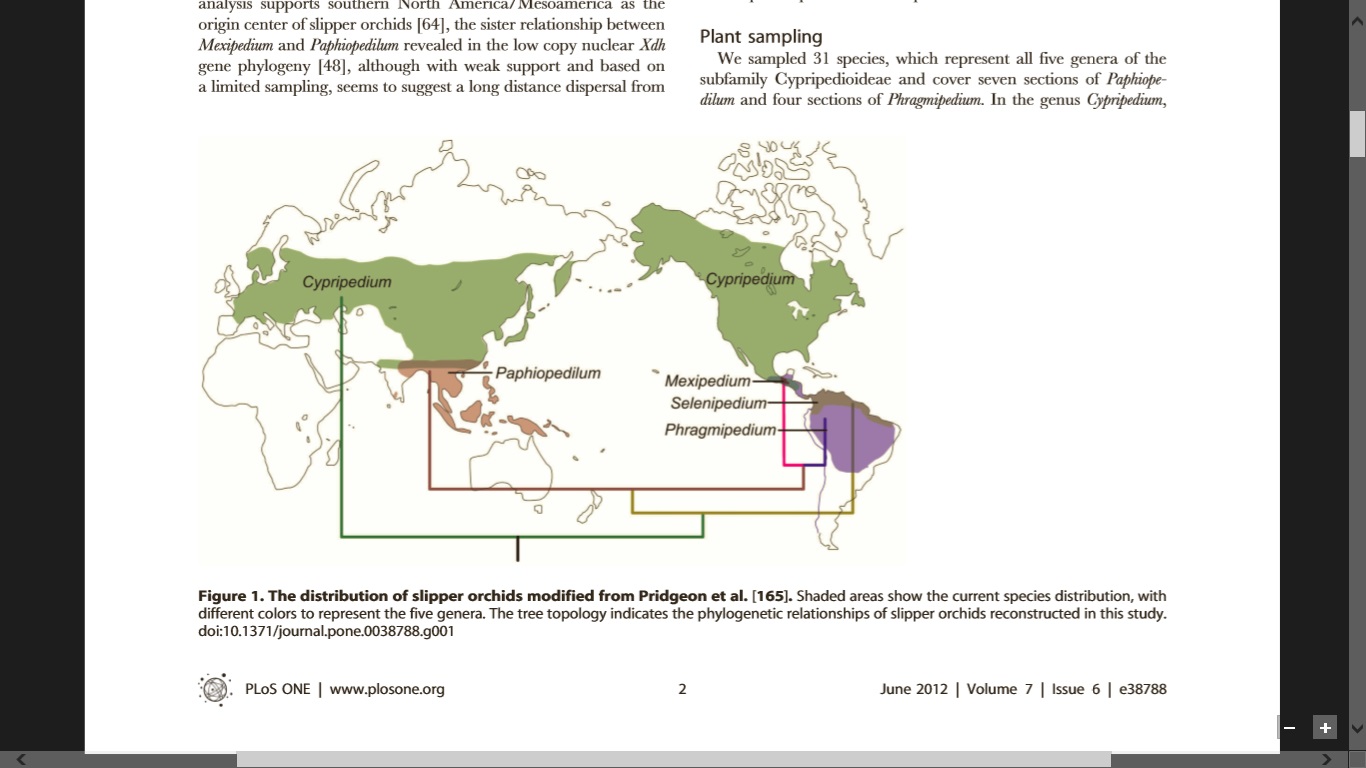
Chloroplast DNA has become widely used in phylogenetic studies, some of the reasons are that (1) it has several copies, (2) no paralogy risks and (3) maternal inheritance (Judd *et al.* 2006). Besides its conserved overall structure, informative, rapidly evolving non-coding plastid regions (NCPs) are found in Orchidaceae (e.g. Yang *et al.* 2013, Jheng *et al*. 2012, Pan *et al*. 2012). Seed plants plastomes have quadripartite structure, composed by two Inverted Repeats (IRs) alternated with a Short Single Copy region and a Long Single Copy region (Figure 3).

* 1. **Markers for Orchid Phylogeny**

Jheng *et al.* (2012) analysed variation among 3 chloroplast whole genome sequences (aka ‘plastomes’), and found that, besides an overall conserved plastome structure, a diversity of phylogenetically informative sequences. Variation in the pattern of loss and truncation of genes encoding for *ndh* subunits, variation in indels patterns of *ndh* peseudogenes, variation of IR junctions, Single Nucleotide Polymorphism (SNP) and variation in intergenic spacers are phylogenetically informative and demonstrate different levels of variations.

Yang *et al.* (2013) found significant differences in variation among plastome regions, being the SSC the most variant region, with 3.5% parsimony informative characters, and the IRs the most conserved regions, with 0.9%. The authors suggest the use of 11 intergenes plastid sequences for Orchid Phylogeny: *cem*A-*pet*A, *clp*P-*psb*B, *ndh*F-*rpl*32, *pet*A-*psb*J, *psb*A-*trn*K, *trn*L-*ccs*A, *rpl*32-*trn*L, *trn*E-t*rn*T, *trn*K-*rps*16, *trn*P-*psa*J, *trn*T-*trn*L along with commonly phylogenetic regions *trn*H-*psb*A. Using intergene hotspots they could clarify relations within the genus *Cymbidium* (Epidendroideae). Chang *et al.* (2006) suggested *ndh*A + *rpl*16 introns for orchid phylogenetic reconstructions. Pan *et al.* (2012) suggested the use of the intergene sequences *trn*F-*ndh*J + *trn*H-*psb*A, and the use of intergene regions to clarify phylogenetic relations within Oncidiinae species.

* 1. **Cypripedioideae: Phylogenetic inference using molecular data**

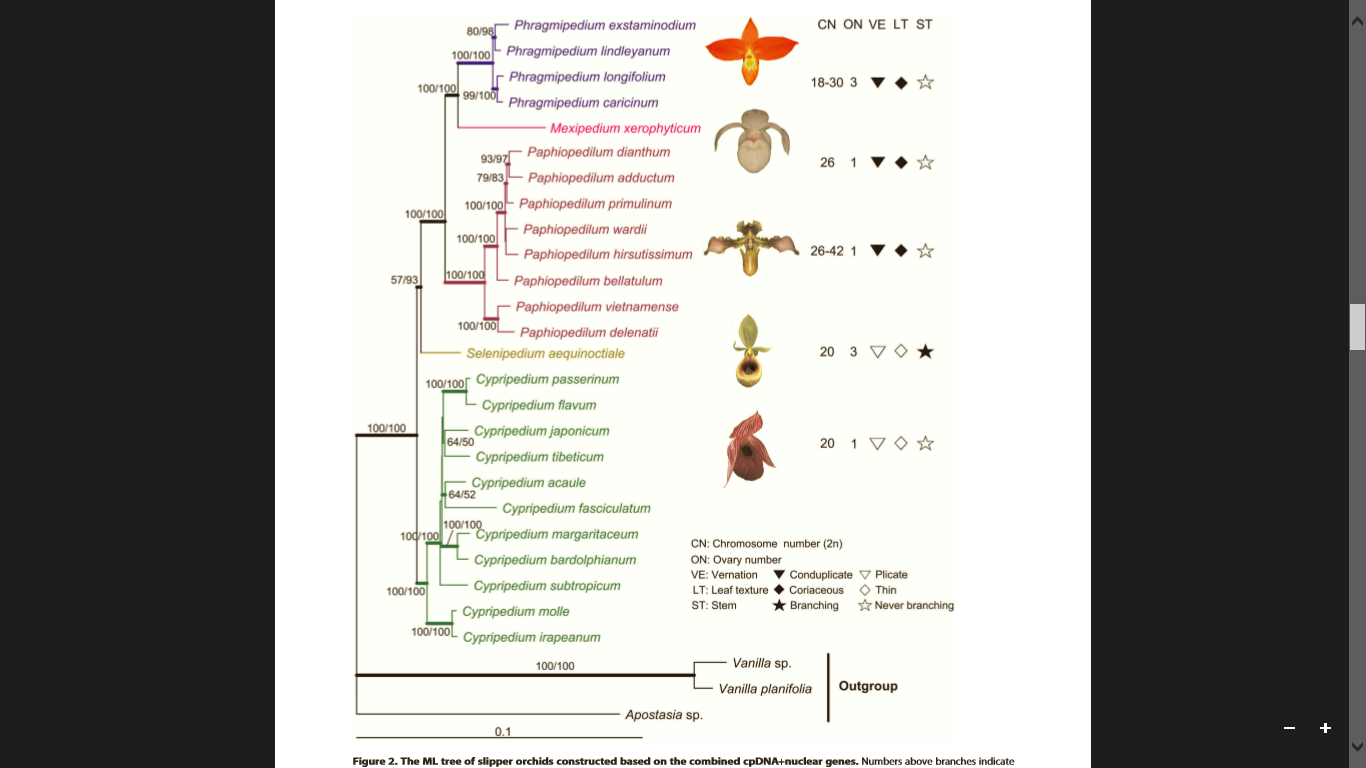
Phylogenetic inference using the alternative markers suggested by Chang *et al.* 2006, Pan *et al.* 2012, Yang *et al*. 2013 and Jheng *et al.* 2012 and newly discovered markers will be conducted in the present study. The efficiency of those markers for Cypripedioideae phylogenetics will be tested. The subfamily contains 5 genera: *Selenipedium, Cypripedium, Mexipedium, Phragmipedium* and *Paphiopedilum*. *Paphiopedilum* is the largest genus in Cypripedioideae*,* with 72 species (Chochai *et al* 2012), followed by *Cypripedium*, with ca. 46 species (Fatihab *et al* 2011). Generic boundaries using morphology have been defined previously using a combination of leaf type, vernation, locule number and placentation (Cox *et al* 1996). Cox *et al* (1996) have delimitated generic boundaries by using the internal spacer (ITS), region of the SSU-5.8S-LSU rDNA, combined with non-molecular data, and using parsimony analysis (including ‘successive weighting’). Generic boundaries using morphology have previously been defined using a combination of leaf type, venation, locule number and placentation (Cox *et al* 1996). Although Cox *et al.* (1996) provided an informative topology in Cypripedioideae, the marker used (ITS rDNA) was not phylogenetically informative above genus level, i.e. it was too slow. In addition, the use of ITS as a marker in phylogenetic studies may provide artefacts (see Álvarez & Wendel 2003), as it may be present in several ‘ribotypes’ in the nuclearorganiser region (NOR), thus the comparison of paralogous genes may occur. The occurrence of homoplasy, compensatory base changes, problems in alignment due to indel accumulation and sequencing errors are the consequences of this erroneous comparison (Álvarez & Wendel 2003).

**Figure 4**: Slipper orchids (Cypripedioideae) distribution among the continents (Guo *et al*. 2012)

Phylogenetic studies within Cypripedioideae have been done using nuclear and plastid genes, and the topology of the subfamily have become relatively well understood (Guo *et al.* 2012). Contradiction is found in some aspects of the topology. For example, using morphological characters (Pfizter 1903) and genetic data (e.g.

plastid gene *rbc*L, Albeit 1994; nuclear nrDNA ITS, Cox *et al.* 1997) have placed the genus *Selenipedium* as sister clade of other slipper orchids. In the other hand, studies that used low copy nuclear genes *xdh* (Cameron 2006), plastid genes *atp*B, *mat*K and *rbc*L (Freudenstein *et al.* 2003), and a combined dataset using plastid sequences (*mat*K, *rbc*L, *rpo*c1, *rpo*c2, *ycf*1, *ycf*2) and low copy nuclear exons (ACO and Leafy) (Guo *et al* 2012). phylogenetic and biogeographical study of slipper orchids was done by Guo *et al.* (2012). The figure 4 shows the distribution of the 5 genera among the continents.

Guo *et al.* (2012) used Bayesian Inference and Maximum Likelihood methods, analysed the phylogenetic and biogeographic relations among 31 species from Cypripedioideae, using concatenated plastid and nuclear genes. In figure 5, the tree obtained by the author is shown. Posterior probabilities (≥0.90) are represented by bold lines and bootstrap values ≥50% are shown above the branches.



**Figure 5**: Phylogeny of slipper orchids, BI and ML using concatenated plastid and nuclear genomes (Guo *et al*. 2012)

The genus *Paphiopedilum* is the most ‘derived’ group in Cypripedioideae. The last monograph was done by Cribb (1998). The genus is fascinating not only because of its ornamental value, but also for the curious characteristic of achlorophyllous guard cells, an abnormal situation compared to the stomata of other plants (Willian *et al.* 1983). *Paphiopedilum barbatum,* as many other species of the genus, is in the Orchid Checklist of endangered species, in this case, sequencing herbarium specimen becomes even more valuable.

As relationships among slipper orchids are relatively well solved, the exploration of alternative cpDNA markers that provided good resolution for other subfamilies will be applied to a combined Cypripedioideae and Epidendroideae phylogenetic inference. The efficiency of those markers will be tested, and if they provide a good resolution, they could be applied to other orchid subfamilies as well, and might even become universal markers for Orchidaceae. In addition, the comparison of complete orchid plastomes among the subfamilies might give insights about the diversification of Cypripedioideae and Epidendroideae.

### **Fungi presence in *P. barbatum* specimens**

Recent findings using IOGA software (dr. Bakker persn. comm.) have detected fungi reads from sequenced herbarium specimens (from SYNTHESYS project), using the nuclear ribosomal Internal Transcribed Spacer (ITS) sequence as bait. Whether my sample also have fungi reads is not known, therefore ITS sequences will be used as bait to infer their presence.

*Tulasnella calospora* (anamorph of *Rhyzoctonia repens*),a ubiquitous mycohrriza, was reported to be associated with roots of *P. barbatum* (specimens of 5 different regions in Malaysia) (Idris 2010). Beyong mycohrrizal species, could also be present sequences from (a) endophytic fungi associated with leave tissue; (b) fungi which have contaminated the voucher specimens during this 46 years guarded in herbarium. Endophtytic fungi are present in the photosynthetic tissues of every tropical plant studied to date, and their diversity is remarkable (Arnold 2008). As DNA extraction from *P. barbatum* was made with leave tissue, it is not expected the presence of *Tulasnella sp*. Nevertheless, the use of *Tulasnella sp.* ITS sequence is expected to be efficient for detecting any species, if it was present in a sufficient amount in the DNA extract.

# Research Questions

* Which plastid markers are phylogenetically informative for analyses between and within two orchid subfamilies (i.e. Cypripedioideae and Epidendroideae)?
* Comparing complete plastomes from different Cypripedioideae (i.e. *Cypripedium, Phragmipedium, Paphiopedilum*) and Epidendroideae genera (i.e. *Bulbophyllum, Elleanthus, Erycina, Oncidium, Phalaenopsis, Rhizanthella, Sobralia*) will give insights about hotspots regions, or rapidly evolving non-coding plastid regions (NCPs).

## 

# Hypothesis

The proposed markers are expected to generate a good resolution for phylogenetic reconstruction between Epidendroideae and Cypripedioideae.

The 5 most phylogenetically informative markers are expected to generate a good resolution for phylogenetic reconstruction among Orchidaceae

The two topologies are expected to be congruent, and the topology generated by the alignment of 14 terminals is expected to be less informative.

# Materials and Methods

## Plastome Reconstruction

### ***Resources Available***

As part of the SYNTHESYS museum DNA project that was finalised January 2012 ([www.synthesys.info](http://www.synthesys.info)), various herbarium genomics data had been generated using Illumina HiSeq NGS technology [(Staats](http://(Staats) & al. PlosONE papers). The Biosystematics Group, Wageningen University (e.g. Bakker & al.) generated data for 94 historical and related fresh samples, including two herbarium specimens of *Paphiopedilum barbatum*. The specimens belong to Naturalis Biodiversity Center, localized in Leiden, under the collection numbers L0717340 and L0717341.

Earlier in this year I spent some time assembling these reads into plastome sequences using a bioinformatics assembly pipeline named IOGA pipeline (developed in the Biosystematics Group by Holmer & Nieuwenhuis, unpublished), however, probably because of the available plastome reference sequences being too distantly related (all from Epidendroideae), my draft-assembly was left with many gaps. Alternatively, failure in obtaining large contigs could have been the result of the very high nuclear/chloroplast genome size ratio in these samples, which were sequenced relatively shallowly (generated from a 24 sample-per-lane Illumina HiSeq run). The size of the nuclear genome of *P. barbatum* is especially big, ~33 Gbp (dr. Bakker pers. comm.). However, it is expected that use of two additional Cypripedioideae reference plastome sequences as scaffold, generated in Kew (*Cypripedium calceolus*) and Florida (*Phragmipedium longifolium*) and made available to dr. Gravendeel (Naturalis, Leiden; pers. comm.) will drastically improve my initial assemblies of these museomes.

As the level of similarity is expected to increase between the references and the target species, the new references should provide a better draft-assembly, with less gaps in between the scaffolds. This provides a good opportunity to reconstruct both *P. barbatum* plastomes sequences, and to explore the use of promising markers for phylogeny reconstruction. Below, more information about both herbarium samples is provided.

**L0717341**

This specimen was collected in March 1970 by C. Davidson, in Terengganu, west Malaysia. The total number of sequences obtained from the Illumina HiSeq2000 run was relatively low: 890,670 reads of ~98 bases length. Although the quality scores from the fastqc report were satisfactory (e.g. no overrepresented sequences), it could not be assembled when using a combined reference of four Epidendroideae orchids. The SYNTHESYS number is 87. Raw data is not yet available online.

#### **L0717340**

This specimen was collected in April 1968 by an anonymous collector in Pennang Hill, Malay Peninsula. The Illumina output included 12,104,266 reads which failed in 7 of the 11 parameters of the fastqc quality report. Regardless of the lower average quality of these reads, the total number of sequences was satisfactory and produced, using the same reference as the previous sample, 46 scaffolds containing a total of 16724 bases. The SYNTHESYS number is 86. Raw data is not yet available online.

### ***Gene Annotation***

Gene annotation will be done using Dual Organellar GenoMe Annotator (DOGMA), as described in IOGA manual, and Geneious.

### ***Submission to GenBank***

The submission of both *P. barbatum* complete plastomes sequences to GenBank will be done using a pipeline developed by dr. Vos (Naturalis, Leiden; pers. comm.).

## Fungi Detection

The Internal Transcribed Spacer (ITS) region of the nuclear ribosomal repeat unit will be used as bait to fish out fungi reads from the raw Illumina readpool. Sequences from the genus *Tulasnella* and *Rhyzotonia* are available on UNITE database (http://unite.ut.ee), a database for molecular identification of fungi (Nilsson *et al.* 2011). ITS sequences from *Tulasnella repens* are available in GenBank as well. The sequences will be used as reference genomes in IOGA pipeline, and, if there are fungi reads, they are expected to be detected. Mycorrhizal endophytic association widely occur in orchidaceae, but its detection is this study is not expected: it occurs mostly in orchid roots, protocorms and plantlets. As only foliar material was sequenced from *P. barbatum* herbarium specimens, either foliar endophytical fungi or fungi contamination of the voucher specimens could be detected. After the obtainment of fungi reads, and after the assembly of those reads, the sequences will be submitted to BLAST search in UNITE database for its identification.

## Phylogenetic methods

### **Whole cpDNA Orchidaceae Alignment: comparative genomic analysis**

Whole plastome sequences within Orchidaceae will be aligned with the aim of calculate similarity among them, methods described in Jheng *et al (*2012). Pairwise distance methods will also be used to detect promising markers. Complete plastomes of species in Arecales, Liliales and Zingiberales will be used as outgroups.

The alignment of complete cpDNA that are available in Leiden (Table 1), comprising orchid cpDNA and related groups among Asparagales could also bring insights about the placement of Cypripedioideae and confirm results of previous work at the family level. However, the low number of terminals on the alignment may not represent Orchid phylogeny with precision. Table 2 shows part of the available complete cpDNA in GenBank that could be incorporated on the alignment. The use of Mesquite for this approach is not enough for this approach. Erik Koenen (pers. comm.) have been using the programming language python to whole chloroplast sequences from a big amount of terminals. Therefore, instead of using mesquite, I will use Linux commands for this alignment next to Geneious.

**Table 1:** Complete chloroplast sequences available in Leiden

|  |  |  |
| --- | --- | --- |
| **Species name** | **Family** | **Subfamily** |
| *Bulbophyllum bicoloratum* | Orchidaceae | Epidendroideae |
| *Bulbophyllum occultum* | Orchidaceae | Epidendroideae |
| *Cypripedium calceolus* | Orchidaceae | Cypripedioidae |
| *Elleanthus sodiroi* | Orchidaceae | Epidendroideae |
| *Masdevallia rosea* | Orchidaceae | Epidendroideae |
| *Oncidium sphacelatum* | Orchidaceae | Epidendroideae |
| *Phragmipedium longifolium* | Orchidaceae | Cypripedioidae |
| *Sobralia bouchei* | Orchidaceae | Epidendroideae |
| *Sobralia callosa* | Orchidaceae | Epidendroideae |
| *Sobralia bouchei* | Orchidaceae | Epidendroideae |

**Table 2:** Complete chloroplast sequences available in GenBank.

|  |  |  |  |
| --- | --- | --- | --- |
| **Species name** | **Acession #** | **Order** | **Family** |
| *Cymbidium mannii* | NC\_021433.1 | Asparagales | Orchidaceae |
| *Cymbidium aloifolium* | NC\_021429.1 | Asparagales | Orchidaceae |
| *Cymbidium tortisepalum* | NC\_021431.1 | Asparagales | Orchidaceae |
| *Cymbidium tracyanum* | NC\_021432.1 | Asparagales | Orchidaceae |
| *Cymbidium sinense* | NC\_021430.1 | Asparagales | Orchidaceae |
| *Phalaenopsis aphrodite* | NC\_007499.1 | Asparagales | Orchidaceae |
| *Phalaenopsis equestris* | NC\_017609.1 | Asparagales | Orchidaceae |
| *Rhizanthella gardneri* | NC\_014874.1 | Asparagales | Orchidaceae |
| *Oncidium Gower Ramsey* | NC\_014056.1 | Asparagales | Orchidaceae |
| *Erycina pusilla* | NC\_018114.1 | Asparagales | Orchidaceae |
| *Veratrum patulum* | NC\_022715.2 | Liliales | Melanthiaceae |
| *Lilium longiflorum* | KC968977.1 | Liliales | Liliaceae |
| *Heliconia collinsiana* | NC\_020362.1 | Zingiberales | Heliconiaceae |
| *Chamaedorea seifrizii* | JX088667.1 | Arecales | Arecaceae |

### **Phylogeny Inference using intergene plastid regions**

Phylogenetic analysis using whole plastome genomes will give insights about marker choice in Orchid phylogeny. An alignment using concatenated sequences will be performed aiming to test the selected sequences. Marker choice will build on suggestions of previous studies: Yang *et al.* (2013), Pan *et al.* (2012) Chang *et al.* (2006) and Gravendeel *et al.* (unpubl.). The considered regions are shown in Table 3, and the marker choice will also depend on the availability of those sequences in GenBank or in the dataset provided by dr. Gravendeel. The number of terminals on the alignment will also depend on the availability of sequences. Marker choice will also build on the results of the previous alignment (iten 4.2.1).

**Table 3:** Phylogenetically informative plastid markers for Orchid Phylogeny in different phylogenetic levels.

|  |  |
| --- | --- |
| Chang *et al*. 2006 | *ndh*A + *rpl1*6 |
| Pan *et al*. 2012 (suggested for intrafamilial inference) | *trnF*-*ndh*J + *trn*H-*psb*A |
| Pan *et al.* 2012 (used in Epidendroideae) | *atp*H-*atp*I, *pet*N-*psb*M, *acc*D-*psa*I, *psb*E-*pet*L, *rps*15-*trn*N |
| Pan *et al*. 2012 (used in subtribe Oncidiinae) | *trn*F-*ndh*J + *trn*H-*psb*A |
| Yang *et al.* 2013 | *cem*A-*pet*A, *clp*P-*psb*B, *ndh*F-*rpl*32, *pet*A-*psb*J, *psb*A-*trn*K, *trn*L-*ccs*A, r*pl*32-trnL, *trn*E-*trn*T, *trn*K-*rps*16, *trn*P-*psa*J, *trn*T-*trn*L + *trn*H-*psb*A |
| Gravendeel *et al.* unpubl. | *ndh*J, *rpl*32-*trn*L, *atp*B-*rbc*L, *rps*11-*rpl*36, *rps*16-*trn*Q, *ndh*J-*ndh*K, *ycf*1, *trn*D-*trn*T, *trn*S-*trn*G, *trn*L-*trn*F, *trn*K-*mat*K, *pet*N-*psb*M, *acc*D-*psa*I,*ndh*C, *trn*C-*trn*D, *pet*B-*pet*D, *ndh*K, *trn*T-*psb*D, *rps*20-*rps*12, *pet*A-*psb*J, *psb*A-*trn*H, *trn*V-*atp*E, *rps*14-*rpl*16, *atp*H-*atp*I, *pet*A-*psb*E, *ycf*3-*rps*4, *atp*A-*atp*F, *psb*A-*ycf*3, *rpo*C1-*rpo*B, *rbc*L, *psb*A, *psb*B |

Gene concatenation will be done using Mesquite. Maximun Likelihood (RAxML) and Bayesian Inference will be performed at the CIPRES platform (phylo.org), and Maximum Parsimony (using Win Paup and TNT) will be performed using various pipelines developed in Leiden.

# Planning

I plan to reconstruct L0717340 and L0717341 plastome sequences using the IOGA pipeline of the Biosystematics group (Wageningen). Genome annotation using DOGMA (IOGA pipeline) will be performed in Wageningen, (in Biosystematics research group) and Geneious analyses will be performed in Leiden (Naturallis). The obtained plastome sequences will be uploaded to NCBI Genbank in Leiden, under the supervision of dr. Vos. Phylogenetic analysis will be conducted in Wageningen under the supervision of dr. Freek Bakker (ML, BI) and in Leiden, under supervision of dr. Gravendeel (MP). The weekly schedule is represented in Table 4.

**Table 4**: Internship weekly schedule

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Literature Study | Writing proposal | Plastome Reconstruction (IOGA) | Genome annotation (WUR) | Genome annotation + GenBank upload (Leiden) | Build Alignment (Leiden) | Phylogenetic analysis (Leiden) | Phylogenetic analysis (WUR BI ML) | Fungi Detection | Writing Report |
| 10-14 Mar | X | X |  |  |  |  |  |  |  |  |
| 17-21 Mar | X | X |  |  |  |  |  |  |  |  |
| 24-28 Mar | X | X |  |  |  |  |  |  |  |  |
| 31 Mar-4 Apr | X | X | X |  |  |  |  |  |  |  |
| 7-11 Apr | X |  | X |  |  |  |  |  |  |  |
| 14-18 Apr | X |  | X |  |  |  |  |  |  |  |
| 21-25 Apr | X |  | X | X |  |  |  |  |  |  |
| 28 Apr- 2 May | X |  |  | X |  |  |  |  |  |  |
| 5-9 May | X |  |  | X |  |  |  |  |  |  |
| 12-16 May | X |  |  |  | X |  |  |  |  |  |
| 19-23 May | X |  |  |  | X | X |  |  |  |  |
| 26-30 May | X |  |  |  |  | X | X |  |  |  |
| 2-6 Jun | X |  |  |  |  |  | X |  |  |  |
| 9-13 Jun | X |  |  |  |  |  | X |  |  |  |
| 16-20 Jun | X |  |  |  |  |  |  | X |  |  |
| 23-27 Jun | X |  |  |  |  |  |  |  | X |  |
| 30 Jun -4 Jul | X |  |  |  |  |  |  |  |  | X |
| 7-11 Jul | X |  |  |  |  |  |  |  |  | X |
| 14-18 Jul | X |  |  |  |  |  |  |  |  | X |
| 28 Jul- 1 Aug | X |  |  |  |  |  |  |  |  | X |
| 28 Jul- 1 Aug | X |  |  |  |  |  |  |  |  | X |
| 4- 8 Aug | X |  |  |  |  |  |  |  |  | X |
| 11-16 Aug | X |  |  |  |  |  |  |  |  | X |

# References

Albert VA (1994) Cladistic relationships of the slipper orchids (Cypripedioi- deae: Orchidaceae) from congruent morphological and molecular data. Lindleyana 9: 115–132.

Álvarez I, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. Molecular Phylogenetics and Evolution 29: 417-434

Arnold E (2008) Endophytic fungi: hidden components of tropical community ecology (Chapter 15 of the book ‘Tropical forest community ecology’ by Carson W P & Schnitzer S A).

Berg C, Higgins WE, Dressler RL, Whitten WM, Soto-Arenas MA, Chase MW (2009) A phylogenetic study of Laeliinae (Orchidaceae) based on combined nuclear and plastid DNA sequences. Annals of Botany 104: 417–430.

Cameron KM, Chase MW, Whitten WM, Kores PJ, Jarrell DC, et al. (1999) A phylogenetic analysis of the Orchidaceae: evidence from rbcL nucleotide sequences. Am J Bot 86: 208–224.

Cameron KM, Chase MW (2000) Nuclear 18S rDNA sequences of Orchidaceae confirm the subfamilial status and circumscription of Vanilloi- deae. In: Wilson KL, Morrison DA, editors. Monocots: Systematics and Evolution. Collingwood: CSIRO publishing 457–464.

Cameron K (2006) A comparison of plastid atpB and rbcL gene sequences for inferring phylogenetic relationships within Orchidaceae. In J. T. Columbus, E. A. Friar, J. M. Porter, L. M. Prince, and M. G. Simpson [eds.]. Monocots: comparative biology and evolution, 2 vols. pp. 447-464.

Cameron K, Arditti J, Kull T (editors) (2007) Orchid Biology Reviews and Perspectives. The New York Botanical Garden Press, Bronx, New York. Vol. 9.

Chang C-C, Lin H-C, Lin I-P, Chow T-Y, Chen H-H, Chen W-H, Cheng C-H, C-Y Lin, Liu S-M, Chang C-C,

Chaw S-M (2006) The Chloroplast Genome of *Phalaenopsis Aphrodite* (Orchidaceae): Comparative Analysis of Evolutionary Rate with that of Grasses and Its Phylogenetic Implications. Mol. Biol. Evol. 23(2):279–291.

Chase MW, Cameron KM, Hills HG, Jarrell D. 1994. Molecular systema- tics of the Orchidaceae and other lilioid monocots. In: Pridgeon A. ed. Proceedings of the 14th World Orchid Conference. London: HMSO, 61–73.

Chase MW, Cameron KM, Barrett RL, Freudenstein JV. 2003. DNA data and Orchidaceae systematics: a new phylogenetic classiﬁcation. In: Dixon KW, Kell SP, Barrett RL, Cribb PJ. eds. Orchid conservation. Kota Kinabalu: Natural History Publications (Borneo), 69–89.

Chase MW, Williams NH, Donisete de Faria A, Neubig KM, Amaral MCE, Whitten WM. 2009. Floral convergence in Oncidiinae (Cymbidieae; Orchidaceae): an expanded concept of Gomesa and a new genus Nohawilliamsia. Annals of Botany 104: 387–402.

Chochai A, Leitch IJ, Ingrouille MJ, Fay MF (2012) Molecular phylogenetics of *Paphiopedilum* (Cypripedioideae; Orchidaceae) based on nuclear ribosomal ITS and plastid sequences. Botanical Journal of the Linnean Society, 2012, 170, 176–196.

Cox AV, Pridgeon AM, Albert VA, Chase MW (1997) Phylogenetics of the slipper orchids (Cypripedioideae, Orchidaceae): nuclear rDNA ITS sequences. Plant Syst Evol 208: 197–223.

Cribb, P. (1997) The genus *Cypripedium*. Timber Press, Portland. 294 pp.

Delannoy E, Fujii S, Colas des Francs-Small C, Brundrett M, Small I (2011) Rampant gene loss in the underground orchid Rhizanthella gardneri highlights evolutionary constraints on plastid genomes. Mol Biol Evol 28: 2077–2086.

Doorduin L, Gravendeel B, Lammers Y, Ariyrek Y, Chin-A-Woeng T, Vrieling K (2011) The Complete Chloroplast Genome of 17 Individuals of Pest Species *Jacobaea vulgaris*: SNPs, Microsatellites and Barcoding Markers for Population and Phylogenetic Studies. DNA Research 18, 93–105.

Dressler R L (1993) Phylogeny and the Classification of the Orchid Family. Cambridge University Press.

Fatihah NHN, Fay MF, Maxted N (2011) Molecular Phylogenetics of *Cypripedium* L. (Cypripedioideae: Orchidaceae) Based on Plastid and Nuclear DNA Sequences. J. Agrobiotech. Vol 2, 2011, p. 35-51.

Fay MF, Chase MW (2009) Orchid biology: from Linnaeus via Darwin to the 21st century. Annals of Botany 104: 359–364

Freudenstein JV, Chase MW (2001) Analysis of mitochondrial nad1b-c intron sequences in Orchidaceae: utility and coding of length-change characters. Syst Bot 26: 643–657.

Góniak M, Paun O, Chase MW (2010) Phylogenetic relationships within Orchidaceae based on a low-copy nuclear coding gene, Xdh: congruence with organellar and nuclear ribosomal DNA results. Mol Phylogenet Evol 56: 784–795.

Govaerts R, Pfahl J, Campacci MA, Holland Baptista D, Tigges H, Shaw J, Cribb P, George A, Kreuz K, Wood J (2011) World checklist of Orchidaceae. The Board of Trustees of the Royal Botanic Gardens, Kew. http://www.kew.org/wcsp/. Accessed 24 march 2013

Gravendeel B, Smithson A, Slik FTW, Schuiteman (2004) Epiphytism and Pollinator Specialization: Drivers for Orchid Diversity? Philosophical Transactions: Biological Sciences, Vol. 359, No. 1450, Plant Phylogeny and the Origin of Major Biomes (Oct. 29, 2004), pp. 1523-1535

Guo Y-Y, Luo Y-B, Liu Z-J, Wang X-Q (2012) Evolution and Biogeography of the Slipper Orchids: Eocene Vicariance of the Conduplicate Genera in the Old and New World Tropics. PLoS ONE 7(6): e38788.

Holmer R, Nieuwenhuis M (unpublished) Iterative Organelle Genome Assembly (IOGA).

Hsu C-C, Chung Y-L, Chen T-C, Lee Y-L, Kuo Y-T, Tsai W-C, Hsiao Y-Y, Chen Y-W, Wu W-L, Chen H-H (2011) BMC Plant Biology11:3.

Idris N A (2010) Identification of a mycorrhiza isolate from Paphiopedilum barbatum (LINDL.) Pfitzer from ex situ location and determination of its growth-enhancing capabilities. Master Thesis submitted to the school of Gradute Studies, University Putra Malaysia.

Jheng C-F, Chen T-C, Lin J-Y, Chen T-C, Wu W-L, Chang C-C (2012) The comparative chloroplast genomic analysis of photosynthetic orchids and developing DNA markers to distinguish *Phalaenopsis* orchids. Plant Science 190: 62-73.

Judd WS, Campbell CS, Kellogg EA, Stevens PF, Donoghue MJ (2002) Plant Systematics: A Phylogenetic Approach, Ed. 2. Sinauer, Sunderland, Mass.

Metzker ML (2010) Sequencing technologies - the next generation. Nature reviews, Genetics Vol 11: 31 – 46.

Neyland R, Urbatsch LE (1996) Phylogeny of Subfamily Epidendroideae (Orchidaceae) Inferred from ndhF Chloroplast Gene Sequences. American Journal of Botany, Vol. 83, No. 9 (Sep., 1996), pp. 1195-1206.

Nilsson R H, Abarenkov K, Larsson K H, Koljalg U (2011) Molecular Identification of Fungi: Rationale, Philosophical Concerns, and the UNITE Database. The Open Applied Informatics Journal, 2011, 5, (Suppl 1-M9) 81-86

Pan I-C, Liao D-C, Wu F-H, Daniell H, Singh ND, et al. (2012) Complete Chloroplast Genome Sequence of an

Orchid Model Plant Candidate*: Erycina pusilla* Apply in Tropical Oncidium Breeding. PLoS ONE 7(4): e34738. doi:10.1371

Pridgeon AM, Cribb PJ, Chase MW, Rasmussen FN. (1999) Genera Orchidacearum. Volume 1. General introduction, Apostasioideae, Cypripedioideae. Oxford: Oxford University Press.

Pridgeon AM, Cribb PJ, Chase MW, Rasmussen FN. (2001) Genera Orchidacearum. Volume 2. Orchidoideae (Part 1). Oxford: Oxford University Press.

Pridgeon AM, Cribb PJ, Chase MW, Rasmussen FN. (2003) Genera Orchidacearum. Volume 3. Orchidoideae (Part 2), Vanilloideae. Oxford: Oxford University Press.

Pridgeon AM, Cribb PJ, Chase MW, Rasmussen FN. (2005) Genera Orchidacearum. Volume 4. Epidendroideae (Part 1). Oxford: Oxford University Press.

Pridgeon AM, Cribb PJ, Chase MW, Rasmussen FN. (2009) Genera Orchidacearum Volume 5. Epidendroideae (Part II). Oxford: Oxford University Press. In press. Genera Orchidacearum

Pridgeon AM, Cribb PJ, Chase MW, Rasmussen FN. (2014) Genera Orchidacearum Volume 6. Epidendroideae (Part III). Oxford: Oxford University Press. In press. Genera Orchidacearum

Pfitzer EHH (1903) Orchidaceae-Pleonandrae. In: Engler A, ed. Das Pflanzenreich. Leipzig: Engelmann. pp 1–132.

Shaw J, Lickey EB, Schiling EE, Small RL (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in Angiosperms: the tortoise and the Hare III. American Journal of Botany 94(3): 275-288

Simpson MG (2010) Plant Systematics 2ed.

Singer RB, Gravendeel B, Cross H, Ramirez SR (2008) The use of orhid pollinia or pollinaria for taxonomic identification. Selbyana 29 (1): 6-9.

Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker JB, Moore MJ, Carlsward BS, Bell CD, Latvis M, Crawley S, Black C, Diouf D, Xi D, Rushworth CA, Gitzendanner MA, Sytsma KJ, Qiu YL, Hilu KW, Davis CC, Sanderson MJ, Beaman RS, Olmstead RG, Judd WS, Donoghue MJ, Soltis PS (2011) Angiosnperm phylogeny: 17 genes, 640 taxa. American Journal of Botany 98 (4): 704-730

Staats M, Cuenca A, Richardson JE, Vrielink-van Ginkel R, Petersen G, et al. (2011) DNA Damage in Plant Herbarium Tissue. PLoS ONE 6(12): e28448.

Staats M, Erkens RHJ, van de Vossenberg B, Wieringa JJ, Kraaijeveld K, et al. (2013) Genomic Treasure Troves: Complete Genome Sequencing of Herbarium and Insect Museum Specimens. PLoS ONE 8(7): e69189. doi:10.1371/journal.pone.0069189

Stevens, P. F. (2001 onwards). Angiosperm Phylogeny Website. Version 12, July 2012 [and more or less continuously updated since]

Williams WE, Grivet C, Zeiger E (1983) Gas Exchange in *Paphiopedilum.* Plant Physiol. 72: 906-908

Wu F-H, Chan M-T, Liao D-C, Hsu C-T, Lee Y-W, Daniell H, Duvall MR, Lin C-S (2010)Complete chloroplast genome of Oncidium Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biology 10:68

Yang J-B, Tang M, Li H-T, Zhang Z-R, Li D-Z (2013) Complete chloroplast genome of the genus Cymbidium: lights into the specied identification, phylogenetic implications and population genetic analyses. *BMC Evolutionary Biology* 13:84