UNIVERSITY OF GOTHENBURG

Phylogenetic Inference and Allopolyploid Speciation

A Study of *Silene* section *Physolychnis*

Anna Petri
2012

Department of Biological and Environmental Sciences
Faculty of Science
University of Gothenburg

ABSTRACT

The major theme of this thesis is allopolyploidization, with focus on *Silene* section *Physolychnis*. The evolutionary history of an allopolyploid species can be established by inferring the sister relationships of the homoeologous sequences from low copy nuclear markers. Sequence specific primers are here used to recover co-amplified genetic variants, which is an efficient alternative to the commonly used bacterial cloning.

A phylogenetic overview of *Physolychnis* is presented, and two major clades within the section are defined: the Asian/American clade and the Arctic/Siberian *S. ajanensis* group. Previously unknown ploidy levels of several taxa are inferred from sequence data, based on the number of monophyletic sequence-clusters recovered per species. Several allopolyploid taxa stemming from crosses between the two major clades are identified. Of these, *S. sachalinensis* and *S. involucrata* have indistinguishable origins, although they are geographically and morphologically distinct. Certain taxa within *Physolychnis* exhibit an extra copy of the nuclear gene *RPA2*, originating from a distantly related lineage within section *Auriculatae*. This is in strong conflict with the species relationships (inferred in several previous studies), and is best explained by introgression between lineages belonging to the two *Silene* subgenera. This finding of gene flow between such long diverged lineages may represent one of the most extreme cases presented yet.

In order to investigate whether *S. sachalinensis* and the three subspecies of *S. involucrata* originate from one or several hybridization events, transcriptome data from three distantly related *Silene* species is used to design a set of primers for 27 loci not previously sequenced in this genus. Together with five other loci, these are amplified from 43 specimens from within the parent / allopolyploid species complex, and sequenced using 454 amplicon sequencing. This method is particularly well suited for a systematic project involving allopolyploid taxa, since sequence variants are separated on the sequencing plate, and due to the possibility of multiplexing a large number of loci and individuals. Here, a two-step PCR approach is taken to attach 10 bp barcodes to the individual amplicons. This approach is time and cost efficient, but may lead to large amounts of recombinant sequences during the second PCR amplification.

Species tree inference is complicated by the presence of homoeologous gene copies within a single species. A phylogenetic tree in which identical taxon labels occur in more than one monophyletic position is here defined as a *multiply labeled tree*. Although several gene-trees-to-species-tree methods exist, the first consensus method that combines several multiply labeled gene trees into a multiply labeled genome tree is presented here. The genome tree can subsequently be folded into a species network, which describes the evolution of allopolyploid taxa.

Keywords: allopolyploidization, hybridization, next-generation sequencing, *Physolychnis*, sequence specific primers, *Silene*, species networks, transcriptomics

LIST OF PAPERS

This thesis is based on the following papers, which are referred to by their roman numericals in the text:

Paper I.
Lott M, Spillner A, Huber K, Petri A, Oxelman B, Moulton M. 2009. Inferring polyploid phylogenies from multiply-labeled gene trees. *BMC Evol. Biol*. 9: 216.

Paper II.
Petri A, Oxelman B. 2011. Phylogenetic relationships within *Silene* (Caryophyllaceae) section *Physolychnis*. *Taxon* 60(4):953–968

Paper III.
Scheen A-C, Pfeil BE, Petri A, Heidari N, Nylinder S, Oxelman B. 2012. Use of allele-specific sequencing primers is an efficient alternative to PCR subcloning of low-copy nuclear genes. *Mol. Ecol. Res.* 12:128–135.

Paper IV.
Bertrand YJK, Petri A, Scheen A-C, Töpel M, Oxelman B. *De novo* transcriptome assembly, annotation, and identification of low-copy number genes in the flowering plant genus *Silene* (Caryophyllaceae). Manuscript.

Paper V.
Petri A, Pfeil BE, Oxelman B. Introgressive hybridization between anciently diverged lineages of *Silene* (Caryophyllaceae). Submitted.

Paper VI.
Petri A, Pfeil BE, Tooming-Klunderud A, Pirani A, Oxelman B. Sequencing of multiple new loci from an allotetraploid species complex and its progenitor lineages in *Silene* (Caryophyllaceae). Manuscript.

Authors' contributions:
I. All authors contributed to the ideas and the development of the algorithm which was implemented by ML. AP and BO provided the biological data set and ensured the biological relevance of the paper. Every author contributed to the writing of the paper.
II. AP was responsible for all parts of the study, BO contributed to the design of the study and the writing of the paper.
III. ACS had main responsibility for the project, all authors contributed with ideas and writing of the paper.
IV. AP, ACS, and BO planned and designed the project. AP made initial contig assemblies and run the GeMprospector analysis, YB did the final filtering, assembly, and annotations, MT contributed with bioinformatics and hardware knowledge. All authors contributed to the writing of the manuscript.
V. AP did the initial discovery during her Master's thesis and had the main responsibility for all parts of the study. All authors contributed to the design and planning of the study, and to the writing of the manuscript.
VI. APE was responsible for all parts of the study. APE, BEP, and BO designed and planned the study, ATK contributed with expertise on the 454 system, API contributed to the lab work. All authors contributed to the writing of the manuscript.

TABLE OF CONTENTS

SVENSK SAMMANFATTNING

Jag blev en gång ombedd att beskriva min forskning med fyra ord som alla kan förstå. "Växters utveckling bakåt i tiden" sa jag då, för "i" är ju inte ett ord, utan en bokstav.

Grunden till **systematiken** handlar om att utreda släktskap mellan arter. Som vetenskap är den gammal, betydligt äldre än 300 år - tiden för Carl von Linnés välkända botaniska arbete. Den moderna systematiska forskningen handlar mindre om utseende, idag är det i stället insidan som är viktig. Morfologiska studier spelar fortfarande en viktig roll, men tack vare möjligheten att analysera kombinationen av de fyra baserna ATGC (vilka bygger upp arvsmassan hos allt liv på jorden), kan systematisk forskning expanderas till betydligt bredare evolutionär forskning. Utifrån långa sekvenser av DNA, var och en med en given kombination av de fyra baserna, bygger man **fylogenetiska träd**, och kan tack vare dem studera evolutionen miljontals år tillbaka i tiden.

Ta ett exempel man kan höra om på nyheterna: "Det finns spår av Neandertal-DNA i nutida människor. Det betyder sannolikt att människan som art inte varit så isolerad som man tidigare kanske velat tro, utan att det föddes en del barn med en förälder av varje art." Just den typen av frågeställningar har jag ägnat mig åt under arbetet med den här avhandlingen, men i stället för människor och Neandertalare har jag studerat Glimmar och Bläror. Det finns många sätt att upptäcka att det finns spår av DNA från en art i en annan arts DNA, men just systematiken tillhandahåller de kanske mest effektiva redskapen för att ta reda på vad det beror på. Det skulle ju kunna vara så att de båda delat en viss mängd genetisk variation i en avlägsen forntid. Men genom att konstruera ett fylogenetiskt träd och jämföra förgreningstider i trädet, kan allt utom just hybridisering uteslutas.

För den som studerar hybridisering är växter betydligt mer intressanta än djur. Djur, i alla fall de som är stora och som de flesta bryr sig om, håller sig för det mesta inom ramarna för sin art. Vissa korsar sig ibland, men avkomman är ofta steril och kan inte ge upphov till någon ny art (mulåsna är ett välkänt exempel). Växter däremot bryr sig inte lika mycket om gränser, utan korsas ofta. Om föräldraarterna är så avlägset besläktade att hybriden blir steril finns en ganska enkel lösning: att dubblera hela genomet (den samlade arvsmassan). Svaret till varför det fungerar hos växter är komplicerat, men en del av det ligger i meiosen – celldelningen där könscellerna bildas. Under meiosen måste varje kromosom hitta sin motsvarighet, den som kommer från den andra föräldern, och placera sig sida vid sida med den. Om de två är för olika kan en hel rad fel uppstå, och resultatet blir ofta sterilitet. Om det däremot finns en exakt kopia av varje kromosom, kan i stället *dubbletterna* hitta varandra, och meiosen kan fungera som normalt. En hybrid med dubbel uppsättning kromosomer är **allopolyploid**, där -polyploid betyder att den har fler än två av varje kromosom (två är det vanligaste i naturen, en från varje förälder). Korsningar mellan en allopolyploid och någon av dess föräldraarter är så gott som alltid sterila på grund av de olika kromosomtalen; återigen är det meiosen som är den svaga länken. Därför kan allopolyploidisering sägas innebära momentan artbildning. Allopolyploidi är mycket vanligt hos växter, och inom den grupp jag har studerat (*Silene* sektion *Physolychnis*) finns flera allopolyploida arter.

Släktet *Silene*, med de svenska namnen Glim och Blära, hör till familjen Nejlikväxter (*Caryophyllaceae*). Det är ett stort släkte med ca 600 arter, spritt över hela norra halvklotet. Den största artrikedomen hos *Silene* finns i medelhavsområdet, men sektion *Physolychnis* har sin främsta utbredning i alpina områden i Amerika och Centralasien. Några arter sträcker sig genom Sibirien och Arktis, ända bort till norra Skandinavien. En stor del av mina studier har involverat *Silene ajanensis*-gruppen, en liten och distinkt grupp inom *Physolychnis* som växer i Sibirien och Ryska fjärran östern. De fyra arterna är **diploider**, vilket innebär att de är helt normala och inte har mer än två uppsättningar kromosomer. Det som gör

dem intressanta är att gruppen har givit upphov till ett antal allopolyploida arter genom hybridiseringar med den Arktiska/Centralasiatiska Polarbläran, *Silene uralensis*. Sådan detaljerad kunskap om ursprung gör det möjligt att ingående studera allopolyploidi ur ett systematiskt perspektiv. Till exempel kan man undersöka om en allopolyploid art har uppstått vid flera olika tillfällen, eller huruvida dagens morfologiska variation inom en grupp allopolyploider motsvarar antalet hybridiseringar som givit upphov till den.

Lite mer specifikt: I avhandlingens andra artikel fokuserar jag på sektion *Physolychnis* som helhet, med syftet att identifiera grupper inom sektionen. Det framgår där att två av allopolyploiderna har identiskt ursprung, trots att de både ser olika ut och har olika utbredningsområden: *Silene sachalinensis* växer enbart på ön Sakhalin utanför Ryska stillahavskusten. *Silene involucrata* å sin sida är variabel och kan delas upp i tre underarter, vitt spridda i Arktis (den har det svenska namnet Fjällglim). Det leder till den mycket intressanta frågeställningen om huruvida *Silene sachalinensis* och de tre underarterna av *Silene involucrata* härstammar från varsitt hybridiseringstillfälle, eller om de alla har ett och samma ursprung, från vilket de har utvecklats åt olika håll.

Extra intressant blir *Silene ajanensis*-gruppen av att det finns DNA från sektion *Auriculatae* i den. De grupperns är så avlägset besläktade att en fertil hybrid mellan dem är uppseendeväckande till och med i växtvärlden!

Trots att allopolyploida arter är så vanliga bland växter är det ännu inte enkelt att arbeta med dem. Därför har en stor del av min avhandling fokuserat på arbetsmetoderna, från labb till analys. Det första steget i många systematiska projekt är att ta reda på den artspecifika kombinationen av de fyra baserna inom en begränsad del av genomet. **Sekvensering** är en standardiserad metod, men eftersom allopolyploider har två skilda **sekvensvarianter**, en från varje föräldraart, måste man på något sätt skilja dem åt för att kunna läsa av DNA-sekvensen. Det kan man göra genom att fysiskt separera de olika DNA-molekylerna före sekvensering, eller genom att begränsa sekvenseringen så att den enbart läser av en variant. Jag har provat två olika sätt, varav det ena är relativt nytt och ännu inte har uppnått perfektion. I sista artikeln reflekterar jag kring detta – vad jag gjorde bra, och vad jag tror kunde gjorts bättre.

Även analyssteget är komplicerat när allopolyploida arter finns med. Om man konstruerar ett fylogenetiskt träd baserat på en viss del av genomet, får man reda på den evolutionära historien för just den delen. Inget annat. Om man vill ta reda på *artens* evolutionära historia måste man sammanföra **genträd** från flera olika delar av genomet till ett enda **artträd**. Det är ett aktivt område inom den systematiska forskningen idag, och det finns därför många bra sätt att konstruera artträd. Men ännu finns det få metoder som passar för allopolyploider, vilka dyker upp på olika platser i varje givet genträd – de är ju närmast släkt med alla sina föräldraarter. Därför har jag i första artikeln varit med att utveckla en metod som tar hänsyn till detta, och som i slutändan skapar – inte ett artträd, utan – ett **nätverk**.

Trädliknelsen har som utgångspunkt att arter uppstår enbart genom att delas upp, när de i själva verket även kan uppstå genom korsningar. Med tanke på hur vanliga hybridarter är, och hur många andra processer som verkar åt motsatt håll från den traditionella bilden av *Livets Träd*, skulle det kanske vara bättre att i stället tala om *Livets Nätverk*?

INTRODUCTION

Hybrid speciation

Nature displays a range of barriers to gene flow between differentiated lineages, such as different pollination syndromes, or the inability of pollen from one species to germinate on the stigma of another. These barriers are likely to be adaptational - interspecific hybrids face a range of difficulties and are often sterile. In animals, genic incompatibilities (where the genes originating from the two parents are so divergent that they cannot interact normally) is the most common cause of hybrid sterility. In plants, chromosomal differences (where the homologous chromosomes are so structurally divergent that meiosis is inhibited) more commonly cause hybrid sterility (Rieseberg, 2001; Rieseberg and Willis, 2007). Still, hybrid species are frequent in nature. A partly fertile homoploid hybrid (with one set of chromosomes from each parent) may achieve isolation from its parental lineages by for example chromosomal rearrangements (Rieseberg, 1997; Lai et al., 2005). After generations of isolation, complete fertility may have been restored and a new species has been established.

More commonly than retaining the ploidy level of their progenitors though, plant species of hybrid origin are allopolyploids (Rieseberg and Willis, 2007). These have both their parental genomes in duplicated (*homoeologous*) copies and the chromosomes can pair with their homoeologue during meiosis, thus restoring gametic viability (but see below). A polyploid is reproductively isolated from its diploid progenitors, since hybrids between them have an uneven set of chromosomes (*aneuploidy*). Polyploidization is thus a mode of the otherwise controversial *sympatric speciation*.

Polyploidization

Polyploidy is the condition of having more than two basic chromosome sets, and is commonly found in plants. Several important crop species are of polyploid origin, for example wheat, oat, cotton, coffee and tobacco. Different ways of measuring ploidy levels give different estimates of the fraction of polyploid plant species, but there is no doubt that they constitute a significant proportion. Commonly, 40-70% of all plant species are estimated to be polyploid (Mallet, 2007), but it is likely that all angiosperms have at least one polyploidization in their evolutionary history (e.g., Soltis et al., 2010). Although polyploidy occurs also among animals, it is far more common in the plant kingdom. Plants appear to tolerate the various deleterious effects of polyploidy better (Comai, 2005), and the establishment of a polyploid species is aided by self-fertilization, parthenogenesis, and clonality - traits often found in plants (Mallet, 2007). Polyploidy also occurs in the fungal kingdom, but receives little attention (see Albertin and Marullo, 2012 for a review).

Biologists normally distinguish between *autopolyploidy* and *allopolyploidy*, which can be defined cytologically or phylogenetically. Here, I employ a phylogenetic definition:

- If the parents belong to the same lineage, the polyploid individual is referred to as an *autopolyploid*.

- If the parents belong to the different lineages, the polyploid individual is referred to as an *allopolyploid*.

In other words, autopolyploidy is the doubling of a genome *within* a species, and allopolyploidy is the doubling of a *hybrid* genome.

Polyploidy may arise by various pathways (see, e.g., Mallet, 2007 and Rieseberg and Willis, 2007 for reviews). The most common of these is gametic non-reduction, where a complete set of chromosomes is distributed into a gamete. If an unreduced gamete fuses with a normal (homoploid) gamete, a triploid individual is formed. This is typically sterile, but may occasionally produce unreduced (triploid) gametes, which may form a tetraploid zygote with a

homoploid gamete. This process is referred to as the '*triploid bridge*'. Alternatively, two diploid gametes may fuse and give rise to a tetraploid zygote, or genome doubling can occur in the early zygotic state. In plants, genome doubling in somatic tissue can give rise to diploid gametes, since they have modular growth (Mallet, 2007). After polyploidization, the two chromosome sets often do not remain identical to their progenitors. Instead, they rearrange to a large extent, exchanging, moving and deleting genes (Soltis and Soltis, 1999). Liu and Wendel (2002) hypothesized that these rearrangements may overcome various genic problems that can arise during the merging of two divergent chromosome sets, and thus aid in the establishment of an allopolyploid species.

Advantages and disadvantages of allopolyploidy

Without attempting to provide an exhaustive list of all possible advantageous and deleterious effects of allopolyploidy, I will here mention a few examples.

When two inbred lines of a species are crossed, the progeny is often more vigorous than the parents due to increased heterozygosity (*heterosis*). Allopolyploids (and hybrids in general) are positively affected by the higher levels of genetic variation, but also the higher gene copy number in itself has advantages. Deleterious recessive alleles are masked, even in the gametic stage (Comai, 2005), and recombination between the homoeologous chromosomes, if it occurs, further elevates genetic variation (Ramsey and Schemske, 2002). Phenotypic variation in allopolyploids is gained not only from genic factors, but also from variation in gene dosage, regulation networks, and epigenetic factors (Osborn et al., 2003). Levels of gene expression is not necessarily additive even in newly formed polyploids, and may change rapidly after polyploid formation (Otto, 2003). As a result, allopolyploid species may display phenotypic traits not possessed by their parents (Otto, 2003), and may be more successful in colonizing new ground. They are for example found in high frequencies in arctic regions, where repeated ice ages have left vast areas uninhabited (Brochmann et al., 2004).

On the down-side of polyploidy is the high C-value, which according to Leitch and Bennet (2004) appears to be under negative selection. They report a significant genome downsizing in polyploid plant species, where the C-value is not linearly related to the ploidy level. Another disadvantage (especially in newly formed polyploids) is the formation of multivalents during meiosis, which may lead to genically or chromosomally unbalanced gametes. Even though selection often restores fertility after a number of generations, polyploid species in general have lower fertility than their diploid progenitors, both in terms of gamete production and gamete survival, and progeny viability (Ramsey and Schemske, 2002). Apart from the internal effects, a newly formed allopolyploid may be rare, and thus have difficulties in establishing (unless it is highly clonal or selfing).

In conclusion, while being potentially deleterious to the individual, allopolyploidy provides nature with genetic variation. Mayrose et al. (2011) report slower diversification rates and higher extinction rates among polyploid lineages, but conclude that even though polyploidy is often an 'evolutionary dead end' it, once established, contributes significantly to the evolution of green plants.

THE STUDY GROUP

In this thesis, I focus on a genus within the Carnation family where hybridization and allopolyploidization is common: *Silene* L. section *Physolychnis* (Benth.) Bocquet.

Silene (*Caryophyllaceae*) is a large genus, with a mainly northern distribution. The number of species varies between taxonomic treatments, but some of the latest studies (Morton, 2005; Melzheimer, 1980; Zhou et al., 2001) estimate the number to between 600 and 700 species. Molecular studies by Popp and Oxelman (2004), Popp and Oxelman (2007),

Erixon and Oxelman (2008), Jenkins and Keller (2010), and Rautenberg et al. (2010) support a subdivision of *Silene* (sensu Oxelman et al., 2001) into two major clades of approximately equal size, subgenus *Silene* and subgenus *Behenantha* (Otth) Endl (=*Behen* (Moench) Bunge).

Silene section *Physolychnis* (subgenus *Behenantha*) is distributed across the Russian Far East, the Arctic and North America, and at high altitudes in Central Asia and South America. The section was monographed by Bocquet (1969), who recognized 61 species. He delimited *Physolychnis* by a combination of characters, where most species have small, often dissected, petal limbs, irregular thyrsoid inflorescences, and erect stems. The presence of five carpels is a constant feature of *Physolychnis* sensu Boquet (1969), but molecular studies by Oxelman and Lidén (1995) and Oxelman et al. (1997) have demonstrated a close relationship between 5-carpelled *Physolychnis* taxa and several three-carpelled taxa, hitherto referred to the sections *Occidentales* Chowdhuri, *Chloranthae* Chowdhuri, and *Odontopetalae* Schischk. ex Chowdhuri. Popp and Oxelman (2007) informally delimited section *Physolychnis* as the least inclusive clade including the Eurasian *S. zawadskii* Herbich (classified in sect. *Odontopetalae* by Chowdhuri, 1957) and the type of the section, *S. uralensis* (Rupr.) Bocquet. This circumscription receives support from molecular studies by Popp and Oxelman (2004), Popp and Oxelman (2007), Erixon and Oxelman (2008), Jenkins and Keller (2010), and Rautenberg et al. (2010).

Three of the major components of *Silene* section *Physolychnis*:
Left: *S. ajanensis* (Rgl. & Til.) Vorosch. Photo by Anja Rautenberg, Uppsala botanical garden (plant grown from seeds originating from Magadan, Russia).
Right: *S. uralensis*. Photo by Pieter van West, Baffin Island, Canada.
Middle: *S. tolmatchevii* Bocquet. Photo by L. Kuznetzova, Khabarovsky Kray, Russia.
Read more about these in the text below.

Natural habitat for central Asian *Physolychnis*
China, 5000 m above sea level, Shula gorge at the border between Yunnan and Tibet. *Silene nigrescens* (Edgew.) Majumdar grows in the gravel. Photo by Magnus Lidén.

AIMS

Paper I. A variety of methods exist to infer a species tree from a number of gene trees. None of these, however, are suited for allopolyploid species, which occur in more than one position in a gene tree. In order to infer the parental lineages of a hybrid species during a multi-gene analysis, each gene phylogeny could first be inferred separately in order to establish the parentage of each gene copy within a species. Homoeologous copies from different genes, but corresponding to the same parental lineage, could then be assigned to the same genome in a multi gene analysis. The term *genome* thus replaces the term *species*. However, individual gene trees may not be perfectly bifurcating and well supported in all cases, especially if high polyploids and several parental lineages are involved, which may impose errors on genome tree inference. In this paper, we aim to develop a consensus method which constructs a *multiply labeled genome tree* from a number of *multiply labeled gene trees*, without the constraint of having the homoeologous gene copies assigned to a parental lineage.

Paper II. The genus *Silene* contains mostly diploid and presumably non-hybrid taxa, but section *Physolychnis* is an exception (Popp and Oxelman, 2004; Popp et al., 2005; Popp and Oxelman, 2007). The aim of this paper is to obtain a phylogenetic overview over *Physolychnis*, and to identify hybrid species and their parental lineages within the section. The North American and the high Arctic *Physolychnis* taxa have been studied in Popp et al. (2005) and Popp and Oxelman (2007), but no molecular studies have extensively covered the Russian and Asian *Physolychnis* before.

Paper III. As an alternative to the laborious method of bacterial PCR subcloning, different gene copies may be recovered using sequence specific PCR and/or sequence primers. We investigate in this paper the usefulness of variant specific sequencing primers.

Paper IV. Transcriptome data from *Silene uralensis* (subgenus *Behenantha*) and *S. schafta* Gmel. ex Hohen. (subgenus *Silene*) is generated to obtain exon information. From this, low copy number regions potentially useful for phylogenetic studies of *Silene* are identified.

Paper V. In this paper, we investigate a case of strong conflict between the *Silene* species tree (known from Oxelman et al., 2001; Popp and Oxelman, 2004; Popp and Oxelman, 2007; Erixon and Oxelman, 2008; Frajman et al., 2009; Jenkins and Keller, 2010; Rautenberg et al., 2010; **paper II**) and the phylogeny of a low copy nuclear gene. We evaluate potential sources of this conflict, namely contamination, gene paralogy, lineage sorting, and introgression.

Paper VI. *Silene involucrata* (Cham. & Schltdl.) Bocquet and *S. sachalinensis* F. Schmidt, two morphologically and geographically distinct allopolyploid species within section *Physolychnis*, were in **paper II** shown to have similar origins. The chloroplast and RNA polymerase markers used in Popp et al. (2005) and **paper II** do not provide the necessary information to resolve the relationships between these two species, the three subspecies of *S. involucrata* (as recognized in Checklist of the Pan Arctic Flora; Elven, 2007), and their parental lineages. The aim of this paper is to investigate the potential of 454 (Margulies et al., 2005) amplicon sequencing in obtaining large amounts of sequence data, which is necessary to resolve this question.

METHODS

During the work of this thesis, I have put much emphasis on the practical issues specific to datasets including allopolyploid taxa, from lab to analysis.

Project planning / choice of genetic markers

Chloroplast DNA (cpDNA) and nuclear ribosomal DNA (nrDNA) are commonly used for phylogenetic inference, since these are easy to amplify. However, plastids are maternally inherited with few exceptions, and cpDNA will therefore only trace the maternal lineage of a hybrid species. Nuclear regions are biparentally inherited, but since the nrDNA loci are tandemly repeated and recombining, they are subject to concerted evolution (e.g., Wendel, 1995; Rauscher et al., 2004). In *Silene* section *Physolychnis*, this process appears to eliminate traces of the maternal lineage only (Popp et al., 2005; unpublished data). Thus, low copy nuclear regions are necessary to confidently infer hybrid origins, but these are not as widely used as cpDNA or nrDNA, and only few primers amplifying regions from a specific study group may be at hand. Of the markers available in *Silene*, the RNA polymerase introns used in, e.g., Popp et al. (2005) and **paper II** may be the most useful for studies of allopolyploid speciation in *Physolychnis*. Their homoeologous copies are retained within the allopolyploids, and they do not appear to undergo recombination. The level of variation is moderate, however, and new regions need to be identified in order to resolve species relationships within several subgroups of *Silene*.

Designing and testing new primers can be time consuming if sequence information from the specific study group is not available. But owing to the potential of the pyrosequencing (Ronaghi et al., 1996; 1998) based Next Generation Sequencing (NGS) methods, genome-wide sequence data can be obtained relatively easily and at an affordable cost. Pyrosequencing is fast and efficient, and enables parallel sequencing such that thousands of sequences (or more) can be generated simultaneously. The transcriptome data from **paper**

IV has been used to develop new PCR and sequencing primers in several projects, two of which are included in this thesis (see below).

PCR and sequencing

Ideally, all variants of a genetic region are equally amplified during PCR. However, it is likely that the primers will amplify certain variants more efficiently than others, with the result that some may not be detected. On the other hand, when Sanger sequencing is employed, the signals from all copies amplified during one PCR reaction will be imposed upon each other in the chromatograms. Bacterial cloning of PCR products is the most commonly used method to separate the copies prior to sequencing. Avoiding this labor intensive method, I have instead tried two other approaches. In **paper II**, I use copy-specific sequencing primers, (see also **paper III**). This approach is straight forward and simple, yielding clear and strong sequencing signals (but see discussion below). In **paper VI**, I use 454 (Margulies et. al., 2005) amplicon sequencing, where single molecules are attached to micro beads, which are distributed into wells on a sequencing plate. 454 generates reads long enough to potentially contain phylogenetic information, and is thus well suited for amplicon sequencing. From each bead, one read is obtained, which means that each PCR amplicon molecule generates its own sequence. Since there are ~1.6 million beads on a plate (Margulies et al., 2005), even rare molecules can be detected in a sample. Thus, the problem of multiple copies, but also the problem of preferential PCR amplification, are largely solved. Variants that are so poorly amplified during PCR that their signals cannot be detected using Sanger sequencing can be found among the reads obtained from a 454 run.

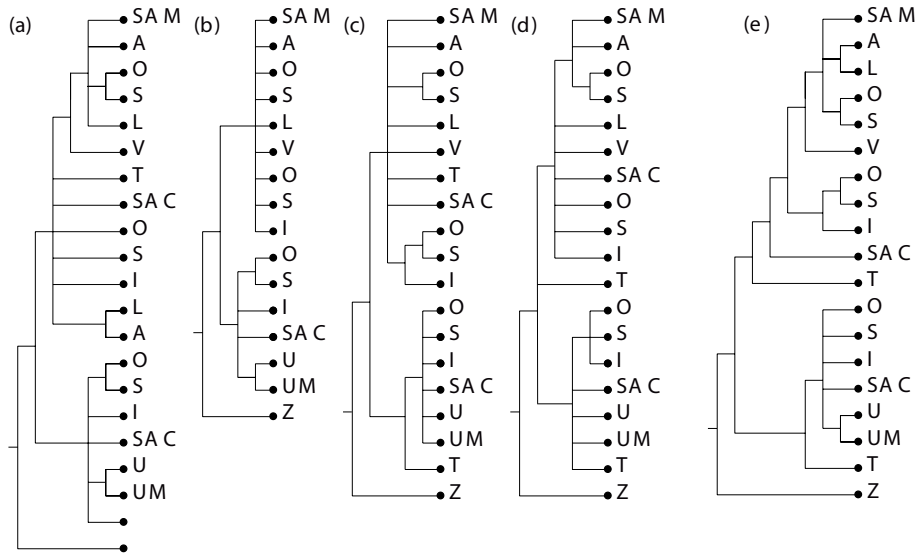
Phylogenetic inference

It is well known that, due to several different processes, the phylogenies of unlinked genetic regions may differ and deviate from the containing species phylogeny. Incomplete lineage sorting and gene paralogy, followed by random loss of one of the copies, leave similar traces in a phylogenetic analysis. Introgression (the transfer of genetic material from one species to another through the formation of a hybrid and the subsequent backcrossing to one of the parental species) and horizontal gene transfer (the transfer of genetic material via a vector) may cause the same topological pattern as lineage sorting, but these can be distinguished if a time scale is inferred from the gene tree. In the presence of allopolyploidization, an extra level of complication is added. If both parental gene copies are retained within the allopolyploid, these are in the simplest case sister to their progenitors and species inference is straight forward. But if gene copy sampling is incomplete, if sister relationships are unsupported, or if the copies are effected by any of the above listed processes, the use of several non-recombining and unlinked low copy nuclear markers is crucial in order to pin-point the processes leading to gene tree / species tree incongruences.

Sophisticated phylogenetic software (such as *BEAST, Heled and Drummond, 2010) can infer species trees from several gene trees while taking lineage sorting into account. These are however not suited for allopolyploid species, since the parentage of each gene copy needs to be defined *a priori* (see above). An alternative is manual inference of hybrid ancestry (see, e.g., Popp et al., 2005 and Maureira-Butler et al., 2008), which is the approach taken in **paper II**.

RESULTS AND DISCUSSION

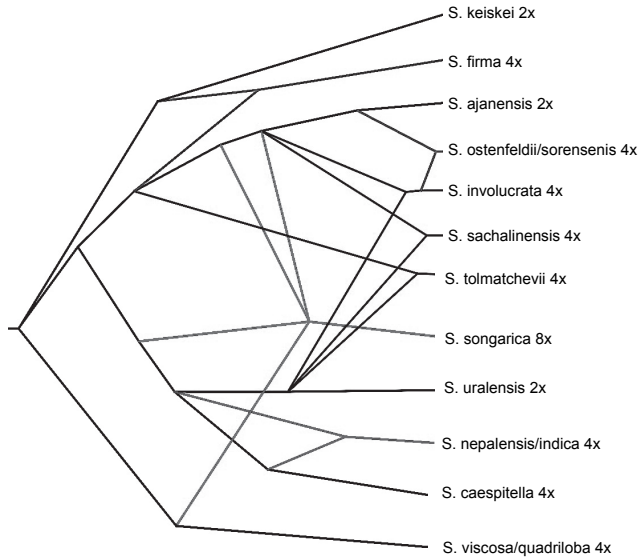
Paper I. The software PADRE identifies clusters in the input gene trees, starting from the smallest and working its way to the root of the tree. It then brings the clusters together into a majority-rule consensus *genome tree*, which can subsequently be folded into a *species network* using an algorithm implemented in the same software (Huber et al., 2006). Hybrid origins of a subset of the taxa from **paper II** are inferred, even in the presence of missing data:



Consensus genome tree from PADRE:
Multiply labeled trees from four low copy nuclear genes from the RNA polymerase gene family (The Arabidopsis Genome Initiative 2000): (a) *RPA2* (b) *RPB2* (c) *RPD2a* (d) *PRD2b*, and (e) the resulting multiply labeled genome tree. Taxon labels are: SAM=*S. samojedora* Lidén & Oxelman (2x), A=*S. ajanensis* (2x), O=*S. ostenfeldii* (A.E. Porsild) J.K.Morton (6x), S=*S. sorensenis* (B.Boivin) Bocquet (6x), L=*S. linnaeana* Vorosch (2x), V=*S. villosula* (Trautv.) V.V. Petrovsky & Elven (2x), T=*S. tolmatchevii* (2x), SAC=*S. sachalinensis* (4x), U=*S. uralensis* (2x), UM=*S. uralensis* (from Mongolia, 2x), Z=*S. zawadskii* (2x, outgroup). The allopolyploid taxa occur at two or three positions in the trees, depending on their ploidy level.
Lott et al. (2009) *BMC Evol. Biol.* 9(216)

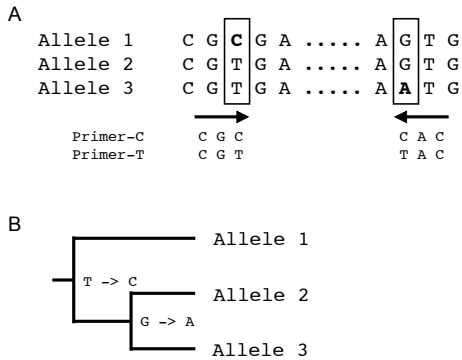
Despite the evolutionary importance of allopolyploidy, this is the first gene trees-to species tree method that is aimed for multiply labeled trees and prior ignorance of parentage. However, it does not implement the multispecies coalescent.

Paper II. A large-scale overview of section *Physolychnis* is presented, and two major clades within the section are defined, the Asian/American clade and the Arctic/Siberian *S. ajanensis* group. Ploidy levels of several taxa, where previously unknown, are inferred from sequence data, based on the number of monophyletic sequence clusters per species. Several hybrid species between the two major clades are identified, some of which appear to have identical origins:



Species network in *Silene* section *Physolychnis*
Allopolyploid origins of some species investigated in **paper II**, as infered from the non-monophyletic positions of their divergent copies of two low copy nuclear genes. Note that the origin of the tetraploid *S. tolmatchevii* is distinct from the putatively common origin of *S. involucrata* and *S. sachalinensis*. Modified from Petri and Oxelman 2011. *Taxon* 60(4): 953–968

The main finding of **paper III** is that variant-specific sequencing primers perform best if kept short. In this way, even primers based on transition polymorphisms may successfully discriminate between sequence variants. The method is efficient and generates clear sequences, but has some limitations. For example, when more than two variants are present in the template, it may not be possible to recover all:



Sequencing with variant specific primers
In the presence of more than two co-amplified genetic variants, sequence recovery may not be successful. In this case, the forward primer matches Allele 2 *and* 3, and the reverse primer matches Allele 1 *and* 2.
Scheen et al. 2012. *Mol. Ecol. Res.* 12(1):128-135.

New specific primers can be constructed at the point of the first polymorphism within one of the initially recovered sequences, but there is a risk that one primer will match also the other

initially recovered sequence (for which it was not aimed). The use of sequence specific primers may thus not be suited for high polyploids.

Paper IV. Annotated transcriptome data from two distantly related *Silene* species are made publically available, along with a large number of identified low copy number genes not previously sequenced in *Silene*.

Paper V. The incongruence between the *RPA2* gene tree and the *Silene* species tree is interpreted as remnant of introgressive hybridization that occurred millions of years after the two involved lineages diverged. Rigorous controls of the lab procedure eliminates contamination, and with the aid of parsimonious reconciliation between gene trees and the species tree, as well as by consideration of clade ages, gene paralogy and lineage sorting can be excluded as alternative hypotheses. This finding of gene flow between such long diverged species is one of the most extreme cases presented yet. There are however indications that the phenomenon may be more common than generally anticipated.

Paper VI. In order to address the number of separate origins of the allotetraploid *S. involucrata* / *S. sachalinensis* group (**paper II**), DNA sequence data from 34 putatively unlinked loci is generated for 43 individuals. Since known markers in *Silene* suitable for such a project are sparse, PCR primers are developed for regions identified in **paper IV**. A two-step PCR approach is taken to tag the amplicons. This is undoubtedly efficient in terms of time and cost of lab work, but in the presence of multiple gene copies, the level of PCR recombination appears to be so high that alternative methods should be considered. Also several other inherent problems of the 454 amplicon sequencing method are discussed. Read length, for example, which limits the possible amplicon length, and the bias toward sequencing of shorter fragments, which obstructs variant detection in longer amplicons.

FUTURE DIRECTIONS

There is a need to develop algorithms that combine a coalescent approach to infer species trees with the approach of working with multiply labeled trees in PADRE. Jones et al. (2012, eprint arXiv:1208.3606 [q-bio.PE]) have recently developed an extension of the *BEAST implementation (Heled and Drummond, 2010), which assigns gene copies to the genomes of an allopolyploid. In this framework, the dataset from **paper VI** can be analyzed, and potentially reveal whether all four allopolyploid taxa have originated from one common hybridization event and subsequently diverged, or whether they descend from separate hybridization events. While gene trees are certainly of great interest, this question requires species phylogenies to be inferred.

Paper V leaves a few questions open. First, self-pollination experiments could reveal whether the different gene copies contained within certain members of *Physolychnis* are alleles or paralogues. Second, using 454 amplicon sequencing, all copies amplified by the different primer combinations used in the "*Multiple primer amplification*" experiment could potentially be recovered. It could thus be more certainly assessed whether one, two, or three copies are retained in the study species. Third, micro-array experiments could reveal whether all copies are expressed, and if so, what their expression patterns are. If they are expressed in different tissues or under different conditions, this could provide a hint to a possible selective advantage of retaining several copies of the *RPA2* gene.

The potential problem of PCR mediated recombination is highlighted in **paper VI**. A large number of putative variants are present in several amplicons, likely resulting from recombination during the second PCR, where initial template concentrations are high. Although reference sequences are not available, it is unlikely that all variants represent real biological variation. However, an empirical survey needs to be performed in order to establish the level of recombination resulting from the second PCR alone. Separation of amplicons from the first

PCR using bacterial cloning, followed by Sanger sequencing would be a straight forward way of assessment. The first-step PCR products should then ideally be the same as those used for the 454 sequencing, or amplification could be repeated under identical conditions. Such an assay would contribute important knowledge about the best approach to barcode allopolyploid specimens and multiple gene copies for a multiplexed 454 amplicon sequencing run.

TACK! THANK YOU!

Från början till slutet, jättetack till min handledare Bengt Oxelman! För att du gav mig möjligheten att börja doktorera, och möjligheten att sluta med det också. Och för allt däremellan. Du har alltid varit en källa till inspirerande idéer, och påminner mig ofta om att det ju är kul med forskning! Du litar till andras förmåga, vi är alla lika viktiga när du bestämmer. Men viktigast av allt, du finns där när det gäller (inte minst under arbetet med avhandlingen), även som moraliskt stöd. Jag behöver väl inte påpeka att avhandlingen inte hade gått att få ihop utan dig. Jag är heller inte säker på att jag hade kunnat slänga ihop den här boken utan min examinator Mari Källersjö. Du styr alltid upp (och muntrar upp), hur hopplöst jag än kan tycka att allt verkar, och får mig att inse att jag nog kan ändå. Ofta tillsammans med min andra examinator Christer Erséus, tack du med! Min andra handledare Magnus Lidén bor en liten bit bort, men finns alltid till hands för att läsa och kommentera, eller bidra med värdefull kunskap om nån växt. Bernard, vilken tur för oss att du kom till just Göteborg! Du ställer alltid upp, och jag har en bestämd känsla av att alla manuskript blir väldigt mycket bättre när du är medförfattare.

Tack igen till Bengt, Mari, Magnus och Bernard för allt ert stöd, och för att ni så flitigt har läst och kommenterat alla olika texter jag har spammat er med under de sista skrivande månaderna.

Alla andra på Botan, ni är så många... Vivian, vad hade labbet varit utan dig? Tack för en nästan oändlig mängd nyttiga labbtips, och för att du alltid kommit med en snabbt hopsnickrad lösning vid krissituation (till exempel i form av en skumgummi-slang). Mats, vad häftigt att du byggde ihop ett kluster! Bara tanken på att köra en enda analys på min lilla dator... There are really so many colleagues that have been to a lot of help in different ways, too many to mention all. A general HURRAY to all PhD students, post docs, assistants, researchers, staff and all that at DPES/BIOENV who could be found in the lunch room for a nice chat, or to help out in some way. Most of all, thanks to* Cajsa, Elisabet, Elisabeth, Filipe, Thomas, Ulrika and Yann for making work not only easier work but also loads and plenty of fun! A very special thanks to Atefeh, the lab work would simply not have been finished in time without you. And for great company and comfort (Hamid too!) during late hours and holidays when there were only us and the lab ghosts in the corridors. Another special thanks to Zeynep, for the supportive soup dinners and late fikas, often including gossip and mutual pep talks. I'm looking forward to see you in your ice princess suit soon again!

Det finns ju systematiska kollegor lite här och var i världen, eller i alla fall Norden. Återigen, alldeles för mycket bra folk för att få plats här... Magnus Popp, min inofficiella handledare, hoppas du är nöjd med vad jag har gjort av ditt gamla arbete! Och Anja, min kompanjon i Långtbortistan och Sileneträsket. Vad skulle jag ha gjort utan dina gedigna kunskaper i ryska, eller din fantastiska databas?

Kanske lika viktigt som alla handledare, examinatorer och kollegor – de som finns i livet utanför bubblan. Att flytta till Göteborg har ju inte bara varit jobb. Alldeles speciellt vill jag skicka en jättekram till mitt kära spelmanslag för en alltid lika skön stämning, trevligt sällskap och grymt bra musik!

Till sist... Tack och åter tusen tack till min fina, kloka mor, och till Sofia, Anna, Maria och alla andra som säger smarta och uppmuntrande saker, om så bara från andra sidan telefonlinjen. Ni har gjort det så mycket lättare att knyta ihop fem års spretigt doktorandarbete till en liten bok.

* In no obvious order

LITERATURE CITED

Albertin W, Marullo P. 2012. Polyploidy in fungi: evolution after whole-genome duplication. *Proc Biol Sci*. 7(1738):2497-509

Bocquet G. 1969. Revisio Physolychnidum (*Silene* sect. *Physolychnis*). *Phanerog. Monogr. 1*. Lehre: J. Cramer.

Brochmann C, Brysting AK, Alsos IG, Borgen L, Grundt HH, Scheen A-C, Elven R. 2004. Polyploidy in arctic plants. *Biol. J. Linn. Soc.* 82: 521–536. doi: 10.1111/j.1095-8312.2004.00s337.x

Chowdhuri PK. 1957. Studies in the genus *Silene. Notes Roy. Bot. Gard. Edinburgh 22:* 221-278.

Comai L. 2005. The advantages and disadvantages of being polyploid. *Nat Rev Genet* 6: 836-46

Elven R. (ed.). 2007 onward. Checklist of the Panarctic Flora (PAF) vascular plants, v.1.0 http://www.binran.ru/infsys/paflist/index.htm (accessed December 2011).

Erixon P, Oxelman B. 2008. Reticulate or tree-like chloroplast DNA evolution in *Sileneae* (Caryophyllaceae)? *Mol. Phylogenet. Evol.* 48: 313--25

Frajman B, Heidari N, Oxelman B. 2009. Phylogenetic relationships of *Atocion* and *Viscaria* (Sileneae, Caryophyllaceae) inferred from chloroplast, nuclear ribosomal, and low-copy gene DNA sequences. *Taxon* 58: 811–824.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol*. 27:570-580.

Huber KT, Oxelman B, Lott M, Moulton V. 2006. Reconstructing the evolutionary history of polyploids from multilabeled trees. *Mol. Biol. Evol.* 23:1784-1791.

Jenkins C, Keller SR. 2010. A phylogenetic comparative study of preadaptation for invasiveness in the genus *Silene* (Caryophyllaceae). *Biol. Invasions*. doi: 10.1007/s10530-010-9907-4

Lai Z, Nakazato T, Salmaso M, Burke JM, Tang S, Knapp SJ, Rieseberg LH. 2005. Extensive chromosomal repatterning and the evolution of sterility barriers in hybrid sunflower species. *Genetics*. 171(1):291-303.

Leitch IJ, Bennett MD. 2004. Genome downsizing in polyploid plants. *Biol. J. Linn. Soc*. 82: 651–663.

Liu B, Wendel JF. 2002. Non-Mendelian Phenomena in Allopolyploid Genome Evolution, *Curr. Genomics* 3:489–506

Mallet J. 2007. Hybrid speciation. *Nature* 446. doi:10.1038/nature05706

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z. 2005: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.

Maureira-Butler IJ, Pfeil BE, Muangprom A, Osborn TC, Doyle JJ. 2008. The Reticulate History of *Medicago* (Fabaceae). *Syst Bio* 57 (3): 466-482.

Mayrose I, Zhan, S. H., Rothfels, C. J., Magnuson-Ford, K., Barker, M. S., Rieseberg, L. H., and Otto, S. P. 2011 *Recently formed polyploid plants diversify at lower rates.* Science 333:1257.

Melzheimer V. 1980 Caryophyllaceae, pp. 353--508. In: Rechinger K.H. (ed.), Flora Iranica, No.163. Akademische Druck-U, Verlagsanstalt, Graz, Austria.

Morton JK. 2005. *Silene*. In: Flora of North America. Editorial Committee, *Flora of North America north of Mexico*, vol. 5. New York: Oxford Univ. Press.

Osborn TC, Chris Pires J, Birchler JA, Auger DL, Chen ZJ, Lee H-S, Comai L, Madlung A, Doerge RW, Colot V, Martienssen RA.2003. Understanding mechanisms of novel gene expression in polyploids. *Trends in Genetics*, 19 (3):141-147.

Otto SP. 2003. In polyploids, one plus one does not equal two. *Trends Ecol. Evol.* 18(9):431-433. doi: 10.1016/S0169-5347(03)00213-1
Oxelman B, Lidén M. 1995. Generic boundaries in the tribe *Sileneae* (Caryophyllaceae) as inferred from nuclear rDNA seguences. *Taxon* 44: 525-542.
Oxelman B, Lidén M, Berglund D. 1997. Chloroplast rps16 intron phylogeny of the tribe *Sileneae* (Caryophyllaceae). *Pl. Syst. Evol.* 206:393-410.
Oxelman B, Lidén M, Rabeler RK, Popp M. 2001. A revised generic classification of the tribe *Sileneae* (Caryophyllaceae). *Nordic J. Bot.* 20:743-748.
Popp M, Oxelman B. 2004. Evolution of a RNA Polymerase gene family in *Silene* (Cayophyllaceae) – Incomplete concerted evolution and topological congruence among paralogues. *Syst. Biol.* 53: 914--932.
Popp M, Erixon P, Eggens F, Oxelman,B. 2005. Origin and evolution of a circumpolar polyploid species complex in *Silene* (Caryophyllaceae) inferred from low copy nuclear RNA polymerase introns, rDNA, and chloroplast DNA. *Syst. Bot.*30: 302-313.
Popp M, Oxelman B. 2007. Origin and evolution of North American polyploid *Silene* (Caryophyllaceae). *Amer. J. Bot.* 94:330-349.
Ramsey J, Schemske DW. 2002. Neopolyploidy in flowering plants. *Annu. Rev. Ecol. Syst.* 33:589–639.
Rauscher JT, Doyle JJ, Brown AHD. 2004. Multiple origins and nrDNA internal transcribed spacer homeologue evolution in the *Glycine tomentella* (Leguminosae) allopolyploid complex. *Genetics*.166:987–998.
Rautenberg A, Hathaway L, Oxelman B, Prentice HC. 2010. Geographic and phylogenetic patterns in *Silene* section *Melandrium* (*Caryophyllaceae*) as inferred from chloroplast and nuclear DNA sequences. *Molec. Phylog. Evol*. 57:978–91.
Rieseberg LH. 1997. Hybrid origins of plant species. *Annu. Rev. Ecol. Syst.* 28:359-89
Rieseberg LH. 2001. Chromosomal rearrangements and speciation. *Trends Ecol Evol.* 16(7):351-358.
Rieseberg LH, Willis JH. 2007. Plant speciation. *Science* 317: 910–914.
Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem*. 242(1):84-89.
Ronaghi M, Uhlén M, Nyrén PA. 1998. sequencing method based on real-time pyrophosphate. *Science* 281, 363–365.
Soltis DE, Buggs RJA, Doyle JJ, Soltis PS. 2010. What we still don't know about polyploidy. *Taxon* 59(5):1387-1403(17)
Soltis DE, Soltis PS. 1999. Polyploidy: recurrent formation and genome evolution. *Trends Ecol. Evo*l. 14(9):348-352.
The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature-London*. 408:796–815.
Wendel JF, Schnabel A, Seelanan T. 1995. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proc. Natl. Acad. Sci*. USA. 92(1): 280–284.
Zhou L, Wu ZY, Lidén M, Oxelman B. 2001. *Silene* (110 spp.). In Wu, Z.Y. & Raven, P.H. (eds) Flora of China 6:66-100. St. Louis: Missouri Botanical Garden.