UNIVERSITY OF GOTHENBURG

Faculty of Science

Department of Biological and Environmental Sciences

Next-generation Molecular Systematics and Evolution

insights into *Medicago*

Filipe de Portugal da Silveira Teixeira de Sousa

2015

Ph.D. Dissertation in Natural Science, specialisation in Biology

ISBN: 978-91-85529-80-3

Printed by Kompendiet, Gothenburg 2015

Abstract

Evolutionary relationships among species have in recent years been inferred using DNA sequences from specific genes and represented by tree diagrams (phylogenies). However, the signal contained in different genes can produce conflicting sets of relationships – i.e., gene tree incongruence. Incongruence among gene trees in *Medicago* L. (Leguminosae) has been observed in several studies but not yet resolved. This thesis aims to re-examine the existing gene phylogenies, generate new data and apply new methods of analysis in order to achieve a deeper understanding of the causes of incongruence among species of *Medicago*.

A comparison and reanalysis of six previously published gene phylogenies for *Medicago* indicates that different biological processes, such as incomplete lineage sorting, paralogy and hybridisation, intervene simultaneously to produce highly incongruent phylogenetic patterns in this plant genus. In order to discern between these different processes, more genomic data is required, and therefore a new approach, using recently developed sequence-capture techniques, is adopted, resulting in the largest, most widely sampled and comparable set of gene phylogenies generated to date for this genus. Identifying causes of incongruence also requires the development a theoretical framework that would allow for the sorting of these data according to different patterns. A model that takes genomic location into account and uses coalescent simulation to compare gene tree topologies is presented. This model is formulated as a flow of tests that culminate in the sorting of data for the inference of species trees, referred to as principal trees when hybridisation is present.

The data produced and the methods developed are used to investigate the phylogeny of two sub-clades within *Medicago*, the *Medicago murex* clade and section Spirocarpos subsection Intertextae. Additionally, coalescent-based species delimitation methods are used to clarify species relationships. A species phylogeny is produced for the *Medicago murex* clade, confirming the separation between *M. murex* and *M. lesinsii*, two species with different karyotype. The inferred species relationships suggest a double event of chromosome speciation in this clade. Two hybridisation events are shown in subsect. Intertextae, both affecting lineages of *Medicago ciliaris*. The principal trees that represent these events are recovered through species tree inference analysis from data sorted by coalescent-based tests. A case of cryptic allopolyploid speciation is discovered in *Medicago prostrata*, a species with known diploid and polyploid karyotypes. Finally, a chloroplast phylogeny of *Medicago* is presented.

This thesis shows that hybridisation has played an important role in the evolutionary history of *Medicago*, both at diploid and polyploid levels. It provides new insights into the evolutionary history of *Medicago* and shows the importance of hybridisation in the evolution of plant species.

Keywords: phylogeny, incomplete lineage sorting, paralogy, hybridisation, *Medicago*

List of Papers

This thesis is based on the following papers, referred to in the text by roman numerals as follows:

I. Patterns of phylogenetic incongruence in *Medicago* L. found among six loci *(manuscript)*

II. Phylogenetic Properties of 50 Nuclear Loci in *Medicago* (Leguminosae) Generated Using Multiplexed Sequence Capture and Next-Generation Sequencing
(PloS one 9.10 (2014): e109704)

III. Using Genomic Location and Coalescent Simulation to identify Paralogy, Hybridisation and Incomplete Lineage Sorting in nuclear loci *(manuscript)*

IV. Cryptic allopolyploidy found in *Medicago prostrata* (Fabaceae) with gene capture and coalescent analysis *(manuscript)*

V. Phylogenetic inference in the presence of hybridisation, paralogy and incomplete lineage sorting: two examples from *Medicago* L. *(manuscript)*

FS was the main responsible for papers I, II, III and IV. The analytical work was shared equally between FS and BEP in paper I and between FS and YJKB in paper III. FS was responsible for lab work and contributed with comments on paper IV.

Table of Contents

Ångrar du din resa?

Nej. Jag ångrar ingenting. Men jag är lite trött.

In *Det sjunde inseglet*, by Ingmar Bergman

Svensk Sammanfattning

Olika processer styr utvecklingen av nya arter. Den viktigaste är kladogenes, vilket inträffar när en art splittras i två. Ofta sker detta via vikarians, som när en ursprungsart gradvis delas upp av geologiska barriärer (t. ex. floder, bergskedjor) eller när en del av arten blir isolerad från resten genom kolonisation av ett nytt område (t. ex. en ö). Andra processer som kan orsaka kladogenes är ekologisk anpassning och enkla reproduktiva barriärer, orsakade ev exempelvis kromosomomlagringar. Det finns emellertid andra mekanismer som komplicerar bilden. En sådan är hybridisering, definierat som reproduktiv korsning mellan två individer från olika arter. Traditionellt anses arter vara reproduktivt isolerade från varandra, eller åtminstone oförmögna att producera fertil avkomma med andra arter. Många exempel motsäger dock denna tes. Ett väl dokumenterat exempel utgörs av allopolyploider, dvs hybrider med fördubblad genomstorlek. Hybridisering kan emellertid förekomma även utan fördubblad genomstorlek , särskilt mellan arter med liknande genomstorlekar och kromosomantal. Detta kallas för homoploid hybridisering och kan endera resultera i att DNA från den ena arten överförs till den andra (introgression) eller att ny art bildas. Det förra är vanligt, medan det senare förefaller vara ganska ovanligt, bara några få väldokumenterade exempel är kända. Hybridisering gör rekonstruktion av den fylogenetiska släktskapen besvärlig, eftersom de flesta fylogenetiska metoder bygger på kladogenes och inte tar hänsyn till hybridisering. Om man till exempel använder sekvenserna från två olika gener för att rekonstruera släktskapen hos en hybridart, kan ju den ena spåra den ena föräldern, och vice versa. De två genträden kommer att se olika ut på grund av detta. Även om hybridisering inte förekommit, kan fylogenetisk släktskapsrekonstruktion vara besvärlig på grund av hur olika alleler (varianter av gener) överförs mellan generationer. När en art delas i två kan de två nya systerarterna bibehålla två eller flera alleler från ursprungsarten. När de två systerarterna ånyo splittras igen, kan de fyra arterna ha olika kombinationer av alleler på ett sätt så att allelerna hos systerarterna inte behöver vara de som är närmast besläktade med varandra. Det här fenomenet kallas djup koalescens, och har som konsekvens att genträden kommer i vissa fall att ha annorlunda topologi jämfört med artträdet. Det finns emellertid matematiska modeller som kan hantera detta, så att artträden kan skattas utifrån vissa antaganden. Förutom hybridisering och djup koalescens behöver man dessutom ta hänsyn till processer som genduplicering och utdöenden när släktskapsförhållanden mellan arter skall rekonstrueras.

I denna avhandling undersöks vilka biologiska processer som orsakar olikheter mellan skattade genträd i alfalfa-släktet (*Medicago*), som till hör växtfamiljen Leguminosae (ärtväxter). Släktet innehåller många arter, av vilka flera är viktiga foderväxter. Tidigare försök att rekonstruera den fylogenetiska släktskapen mellan arterna har resulterat i sinsemellan mycket olika resultat. För att söka förstå oraken till dessa olikheter har vi med hjälp av sekvenser från genomet hos *Medicago* truncatula valt ut ett antal ytterligare gener för analys. Emellertid behövs en effektiv och pålitlig metod för att få fram sekvenser från ett stort antal gener för ett stort antal arter. Vi använde nyligen utvecklade ”sequence-capture”-tekniker, som möjliggör anrikning av hudratals gener från många individer. Därefter använde vi ny ”Next Generation Sequencing” (NGS)-teknik för att få fram DNA-sekvenserna. Från dessa kunde vi rekonstuera ett stort antal genträd för *Medicago*. För att förstå vilka biologiska processer som orsakat skillnaderna mellan genträden utvecklade vi metodologiska modeller som kan hantera de processer (hybridisering, djup koalescens och genduplicering) som orsaker skillnaderna. Med hjälp av våra data och dessa metoder kunde vi belysa släktet *Medicago*’s evolutionära historik. Vi kunde demonstrera existensen av kryptiska allopolyploida linjer, och vi undersökte artbildning hos diploida linjer med hjälp av nyligen utvecklade metoder för artavgränsning. *Medicago* murex bör splittras i två arter med olika kromosomantal. I gruppen *Medicago* sect. Spirocarpos subsect. Intertextae har minst två

hybridiseringstillfällen förekommit. Analyser av kloroplastgenomet visar att traditionell taxonomisk indelning av släktet har mycket lite stöd av molekylära data. Våra resultat visar på nyttan av att använda många gener för att förstå arternas släktskap med varandra inom *Medicago*. När hybridisering har förekommit är koalescent-baserade modeller för att jämföra och sortera genträd till stor hjälp för att spåra föräldra-arternas bidrag till hybriderna. Vi visar på existensen av homoploid hybridisering i *Medicago*, åtminstone i form av introgression, vilket gör de tidigare observerade konflikterna mellan genträden begripliga.

Introduction

The importance of Systematics for understanding Biodiversity

Biodiversity is one of the most important words of our time, as it refers to all life on Earth. It can be defined as "the diversity of life at all levels of biological organisation" (Gaston and Spicer, 2013) and includes anything from a single microbial cell to a biome, such as tropical forests, but is commonly perceived, for simplification, merely as a sum of living species, either globally or in a certain area. Systematics is the field of science that uncovers Biodiversity, allowing it to be further studied and eventually become valued, used and preserved. The product of systematic studies, which can be defined as the cataloguing of life forms and their relationships (Wiens, 2007) can have implications on several human activities such as in agriculture, for example through crop improvement (Gur and Zamir 2004 ; Hajar and Hodgkin, 2007). Understanding phylogenetic relationships also allows for the inference of past geological and climatic conditions and their effect on biodiversity (e.g. Oberprieler 2005), thus providing essential information for the understanding of current and future changes in climate and in the biosphere. Systematics went through a major revolution in recent decades, with the advent of molecular techniques. However, the interpretation of molecular sequence data is not always simple, since speciation, and consequently the evolution of genomes, are shaped by various biological processes.

The problem of phylogenetic incongruence

Homologous molecular characters and evolutionary models

The development of nucleic acid sequencing provided a new source of data that is almost unlimited. The field of systematics has been one of the most benefited from this methodological development, since the inference of evolutionary relationships had always been dependent on morphological and biochemical data, both of which have severe limitations, in terms of quantity and of homology assessment. Phylogenetic inference based on genomic molecular data assumes that inherited ancestral characteristics (nucleotides) undergo evolutionary changes (Swofford et al. 1996) and can thus be compared to estimate relationships. However, comparing nucleotide data has its challenges and requires the appropriate tools (Kumar et al. 2004), as misalignment of sequences may originate non-homologous comparisons, which can mislead phylogenetic estimation (Ogden and Rosenberg 2005). Molecular datasets must also be adequately assessed prior to phylogenetic analyses in order to avoid systematic error, namely though the testing of appropriate substitution models describing parameters related to changes between nucleotides (Swofford et al. 1996).

The coalescent: incomplete lineage sorting

When dealing with molecular phylogenetics it is important to distinguish between gene and species trees. Gene trees are phylogenetic trees inferred from a portion of a genome, such as an entire, or part of, a functional gene. Species trees, on the other hand, represent, ideally, the evolutionary history of whole populations. Gene trees often differ from each other and from the corresponding species tree (Degnan and Rosenberg 2006; Kubatko and Degnan 2007; Liu and Pearl 2007). The processes behind the transmission of genes in sexually reproducing species, namely the separation of homologous chromosomes at meiosis prior to fertilization, enable alleles (variants of the same gene in each locus) to follow independent evolutionary paths within populations. Furthermore, recombination between homologous chromosomes makes alleles from different loci independent from one another. Each individual inherits one allele (at a given locus) from each parent but transmits only one allele to each descendent. With time, one of the alleles within a population tends

to be fixed and passed on to descendent species after a speciation event, the other alleles becoming extinct. However, if a population is large enough, it may sustain the multiple alleles up to the point of speciation. In such cases, the descendant species of such a large population may inherit and fix different alleles. After multiple speciation events, tracing the evolutionary history of one of these alleles reveals a pattern that differs from that of the species evolutionary history. The inheritance of alleles within a population is explained by the Coalescent model (Kingman, 1982), and the multi-species coalescent model (Rannala and Yang, 2003).

Deep coalescence, or incomplete lineage sorting, is a prevalent source of phylogenetic incongruence (Degnan and Rosenberg 2006; Kubatko and Degnan 2007). It occurs when at least two allele lineages are present in a population. After a first speciation event, one lineage is fixed in one of the descendent species, while the other descendent species retains the two alleles (Figure 1). If the species that retained both alleles goes through a second speciation process, and each of the alleles is fixed in one of the descendent species, the alleles of these two sister species will coalesce deeper than the population that originated the two species, i.e., the most recent common ancestor of the two alleles will not be found immediately before the last speciation event, but before at least two speciation events. A molecular analysis of these alleles will reveal allele relationships that are incongruent with other gene phylogenies and with the true species phylogeny (Figure 1). Incomplete lineage sorting (ILS) may pose a complication to species tree inference but can be dealt with, for instance, with the multispecies coalescent model (Rannala and Yang 2003). This model reconciles data from multiple loci to estimate coalescent times, population sizes and species tree topologies, and is widely used for phylogentic reconstruction. Biological processes other than ILS, however, can result in incorrect phylogenetic inference, and require special attention.

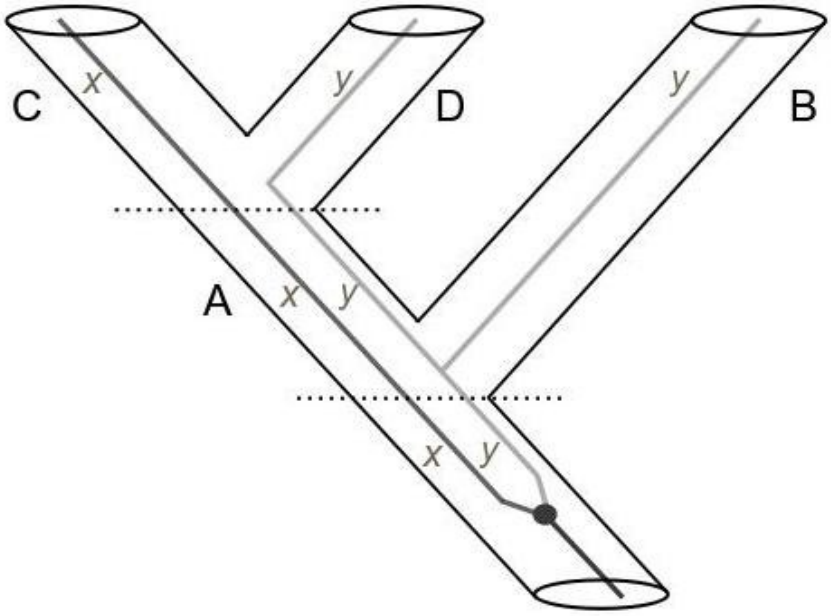

Figure 1: A simplified illustration of incomplete lineage sorting affecting one locus. Speciation events are marked by doted lines. A circle marks the origin of alleles x and y in the initial population. After the speciation event that originates species A and B, both alleles (x and y) are present in species A, but only allele y is kept in species B (allele x is extinct in B). After a second speciation event which splits species A into C and D, species C inherits allele x and species D inherits allele y. The true species phylogeny is (B(C+D)) but sampling this locus would result in a tree with the topology (C(B+D)).

Paralogy

One of these processes is paralogy, which results from gene or whole genome duplication. Gene and genome duplications are common in plants, animals and other organisms (Lynch and Connery 2000; Blanc and Wolfe 2004; Cui et al. 2006). With time, duplicated genes tend to revert to single copy, but duplicates may be retained for extensive periods of time and through many speciation events (Lynch and Connery 2000). The effect of paralogy on gene trees can differ, depending on the age of the duplication and on the duplication-loss dynamics. When genes revert to single copy in most extant species, there may be no evident trace of the duplication event. The differential loss of copies in different species can have dramatic effects on gene trees, because each species retains one or the other copy. If all species retain the same copy of the gene, the corresponding gene topology would be within the expectations of ILS, but if each of the gene copies is retained by different species the resulting gene tree topology may be quite different from what is expected under a scenario of ILS (see Paper III, Figure 1d). Therefore, the identification of paralogous genes is fundamental prior to species tree inference. When more than one gene copy is retained in at least one species, paralogous sequences can often be detected in alignments themselves. When differences are not as obvious, paralogues may be mistaken for orthologous sequences. In this case, when it is impossible to know if sequences are orthologous or paralogous, detection of paralogy becomes difficult, but may be achieved by comparing gene trees, with the use, for example, of tree-distance methods (see Paper III). Currently, the effect of paralogy on phylogenetic inference and the solutions to accommodate its effect remain a challenge in molecular systematics (Rasmussen and Kellis 2012).

Hybridisation

Yet another and highly complex source of phylogenetic conflict is hybridisation (Sang and Zhong 2000). The most common consequence of hybridisation is simple genetic introgression between species, which can occur in different degrees and over more or less extended periods of time. The degree in which it occurs, the duration and direction of the gene flow between the species involved, the size of the affected populations and the gene flow between these and other, non-affected populations, determine how phylogenetic signal will be influenced by introgression. Another less common and less well-documented form of hybridisation is homoploid hybrid speciation. Unlike allopolyploid formation, which arises when hybrid individuals retain two or more entire parental genomes (either by the combination of unreduced gametes or by doubling chromosomes after fertilisation) and often leaves recognisable signs, not only in karyotypes but also at the morphological level (e.g., crop cereals), homoploid hybrid speciation is difficult to detect. The traditional view is that inter-specific homoploid hybrids have poor reproductive fitness, but studies in *Helianthus* (Rieseberg 1991; Rieseberg 1997) show the existence of species that originated from populations of hybrid individuals of two parental species, whose gene flow with the parental populations had a fixed duration in time. The extent to which such species occur is not well known, and becomes an even more difficult matter if hybrid speciation is ancient. Nevertheless, homoploid hybridisation, with subsequent speciation or simply as introgression between diploid species, may provide an explanation for many phylogenetic patterns observed in extant species. The coalescent model for species tree inference is violated by any hybridisation event, since the genome of hybrid individuals contains genes that trace different species (cladogenetic) histories, and hybridisation can be easily mistaken by ILS, but several models have been developed to accommodate this problem (Holland et al. 2008; Joly et al. 2009; Kubatko 2009; Meng and Kubatko 2009).

Species trees versus Species networks

The existence of gene flow between entities defined as species, often resulting in the formation of independent lineages, such as in the case of allopolyploidisation, entails the reconsideration of trees as the best representatives of species phylogenies. The fact that allopolyploids, which originate in hybridisation events, occur frequently, at least within the Angiosperms (Wendel 2000; Petri and Oxelman 2011; Bertrand et al. 2015), is in itself a reason to reject trees as representations of species phylogenies. Even among diploid taxa there may be gene flow, resulting in introgressed genes or even formation of new lineages (Rieseberg 1991). When hybridisation is present, a complete representation of species relationships can only be achieved through a network, such as a splits graph. Networks can be represented by multillabelled trees, in which species affected by hybridisation between two or more lineages are represented by two or more terminals, or labels (e.g. Jones et al, 2013; Bertrand et al. 2015; see Paper I, Figures 7 and 8). Networks can also be decomposed into principal trees (Holland et al. 2008), which reflect all possible placements of hybrid individuals in alternative parental lineages. Each hybridisation event (in the simplest case) is represented by two principal trees. If different independent hybridisation events occur, the number of possible principal trees equals the number of possible combinations for different taxon placement (e.g. four principal trees for two hybridisation events; see Paper V, Figure 2).

Methodological novelties in systematics

Next-generation Sequencing

Recent years have seen a major advance for biological sciences in general, as Sanger sequencing and PCR-based sampling methods are being replaced by Next-Generation Sequencing (NGS) and novel enrichment methods. Next-generation sequencing techniques have increased the capacity of DNA sequencing exponentially (Cronn et al. 2012; Lemmon and Lemmon. 2013) . The most important consequence of that technological shift has perhaps been the increase in the number of sequenced genomes of higher organisms. The availability of near-complete genomic data has opened new possibilities in plant systematics. It is now possible to obtain information about an almost unlimited number of loci. This includes information on gene structure, such as the size and positions of coding (exons) and non-coding (introns, UTRs) segments, on gene synteny (location of genes on the chromosomes), copy number, expression levels, homology with genes in other genomes, and of course, the sequence of the genes. Having access to such information makes it possible to choose phylogenetic markers according to desired properties, for instance, intron number and copy number. Sequencing of a high number of loci for multiple species is enabled by NGS technology, but still requires appropriate enrichment methods for sampling genomes.

Sequence-capture

Phylogenies at high taxonomic levels within the plant Kingdom and within flowering plants have been obtained with considerable success from few molecular markers, but well-supported and conflict-free species-level phylogenies in plants are still a challenge and require the development of additional molecular markers (Hughes et al, 2006). Currently, the two main options for obtaining multiple molecular markers in plant groups are anonymous loci approaches, such as genome skimming (Malé et al. 2014; Bock et al 2014) and RAD-tags (Eaton and Ree 2013; Wang et al. 2013), and target enrichment methods, such as sequence-capture (Lemmon et al. 2012 ; Stull et al. 2013; Carstens et al. 2013; Smith et al 2014). Each approach has its own advantages and can be more suitable than the other according to the sort of biological question that is addressed. For instance, RAD-tags, which are based on sequences adjacent to enzymatic restriction sites, have the

advantage of not requiring reference sequences and produce a very high number of comparable characters distributed quite evenly over genomes, and may be the best option when dealing with population-level questions. Sequence-capture, on the other hand, requires background reference sequences, i.e., either a known genome or transcriptome, but is able to generate longer comparable sequences than RAD-tags. This an advantage if coalescent-based methods, which generally require gene-tree inference (but see Bryant et al. 2012), are chosen to resolve species relationships, as loci chosen a priori are generally longer and usually contain exon and intron sequence, allowing for the inference of more well-supported gene trees.

Bioinformatics – assembly, mapping, allele phasing

Sequencing techniques, as well as compatible enrichment methods, have recently gone through exponential developments (McCormack et al. 2013), and required an equally fast development of analytical tools to process all the data produced. The two approaches to next-generation sequencing data processing are (1) *de-novo* assembly, a complex procedure where individual sequence reads are put together, through bioinformatic algorithms, to produce complete DNA sequences, and (2) sequence mapping, in which sequence reads are instead mapped directly onto a reference sequence and combined to produce the new sequence, homologous to the reference. *De-novo* assembly is the procedure used when sequencing entire genomes, since between organisms there are not only changes in sequences themselves but also changes in gene structure and synteny, i.e., location of genes within a genome, and thus the use of a reference may cause erroneous results. However, when generating sequences for multiple loci in a group of closely related organisms, as is often the case with sequence-capture methods, choosing a mapping approach is convenient, as it tends to be less time-consuming. If the reference used for sequence capture is a genome, then sequence mapping can be done directly onto the sequences of target loci. If, on the other hand, the reference was a transcriptome, and intron sequence data is also desired, then a new reference must be produced, through de-novo assembly of the sequences of one or more individuals (to introduce the intron sequences into the new reference), in order to proceed with read mapping.

Medicago

etymology, higher taxa, species diversity and evolution

Medicago is a genus in family Leguminosae (=Fabaceae), subfamily Papilionoideae, tribe Trifoliae, subtribe Trigonellinae (Lewis et al. 2005), and comprises 87 species (Small 2011). The origin of the name derives from the Latin word *medica*, referring to the ancient empire of Media (Mikaili and Shayegh 2011) which extended from Anatolia to the valley of the Indus. The taxonomy of the genus has been vastly studied, namely by authors such as Širjaev, Lesins and Lesins, Vassilczenko, C.C. Heyn and, more recently, by E. Small, whose monograph on alfalfa contains the taxonomic treatment used as the basis for this thesis. Species of *Medicago* include perennial herbs (e.g. *M. sativa*) and woody sub-shrubs (*M. arborea*) as well as annual herbs (*M. truncatula*). Genera in subtribe Trigonellineae, i.e. *Medicago, Trigonella and Melilotus* (which is nested in *Trigonella*), can be readily distinguished from the remaining members of tribe Trifoliae by the central leaflet of trifoliate leaves, which in members of Trigonellinae has a longer stalk than the other leaflets, opposed to evenly stalked leaflets in subtribe Trifolinae. However, distinguishing genera within Trigonellinae, and especially separating *Medicago* and *Trigonella*, has been a controversial subject. Early attempts were made to separate both genera on the basis of simple morphological characters, such as the pulvini in cotyledons and the typical coiled pods of *Medicago* species, but these characters are not present in all members of the genus.

Small (2011) proposed the the separation of *Medicago* from allied genera on the basis of its explosive pollination mechanism, which is triggered when the pollinator comes into contact with the horn of wing petals of the flowers, causing the explosive release of the staminal column and pistil, which are then pressed against the abdomen of the pollinator. The horn of the wing petals and other floral characters are involved in this unique mechanism. Thus, *Medicago* can be separated from *Trigonella* for having wing petals with large horns, keel petals with small apical notch and stigma less than half the total length of the pistil. *Medicago* also has a mushroom-like stigma and an arched staminal column with unevenly sized filaments. *Medicago* is also distinct in having medicagenic acid glycosides, which are absent in closely related genera.

Importance as crop

Medicago is currently one of the most important crops in the world (Small 2011) but has been widely cultivated for millenia for its use as fodder, forage and for human consumption. The most widely cultivated species, *M. sativa* (alfalfa), originates in central Asia, and its cultivation is believed to have begun eight to nine thousand years ago (Ivanov 1977) in the regions around the southern part of the Caspian Sea (Quiros 1988) from where it spread to neighbouring regions such as China (Mikaili and Shayegh 2011), and accompanied the domestication of the horse (Small 1995). The main use of alfalfa today is animal nutrition, primarily as fodder for dairy cows but also for other farm animals, including poultry. It is also an important honey crop in North America. Besides being harvested as fodder, *Medicago* is also cultivated as permanent pasture (forage), as green manure, as ground cover to prevent soil erosion, as companion to other crops and for habitat enrichment (Bauchan and Greene 2000). Direct human consumption of alfalfa is mostly restricted to fresh seed sprouts, but due to its high protein level there is interest in adding it to human diet in indirect ways (e.g. Huyghe et al. 2007). Its also has potential for the production of biopharmaceutical products, bioplastics, cellulose and biofuel (Small 2011).

Medicago as a model organism

Annual species of *Medicago* have a fast life-cycle, which can be as short as three months from germination to frutification, making them ideal candidates for model organisms. Together with the economic importance of *Medicago*, this was enough to justifiy the choice of *Medicago truncatula* as one of the first plants to have its entire genome sequenced and annotated. The first annotated version of the genome of *M. truncatula* was made available in 2010 (Young et al 2011).
Besides producing a near-complete sequence of all eight chromosomes of *M. truncatula*, the sequencing effort was followed by an exhaustive annotation that identified genes and predicted heir function, identified the plant tissues where genes are expressed and the degree of the expression, indicated putative homologues of each gene in other plant genomes (e.g. *Populus*, *Lotus*), and provided information on gene structure, namely the position of exons (coding sequence), introns non-coding sequence) and UTRs (untranslated regions). Such detailed information about a genome is useful to many research fields, including systematics. Having access to this kind of information is an ideal situation for anyone intending to sequence high number of loci for phylogenetic studies using NGS techniques, since loci can be chosen according to the desired properties.

Phylogeny

The evolutionary history of *Medicago* is currently unsolved. All phylogenies produced thus far (Downie et al., 1998; Bena et al., 1998; Bena et al., 2001, Maureira-Butler et al., 2008, Steele et al., 2010, Yoder et al. 2013) are highly incongruent and have not resulted in an estimation of a species tree or species network for the genus. It is clear that more molecular data must be collected in order

to clarify species relationships and explore the causes of the observed phylogenetic incongruence. It is also necessary to analyse the problem in parts. For instance, polyploid species, which may originate through allopolyploidy, need to be carefully analysed, rather than being incorporated in genus-wide phylogenies. Also, patterns of incongruence that do not resemble incomplete lineage sorting need to be carefully investigated, as biological processes such as hybridisation or introgression and paralogy may be driving the observed patterns. Hybridisation is , in fact, known to occur spontaneously in *Medicago* (e.g. Şakiroğlu et al. 2010; Small 2011).

Objectives

The objective of this thesis is to provide insights into the evolutionary history of *Medicago* through a better understanding of the sources of phylogenetic incongruence in the genus. The first step to achieve this goal is to explore the current knowledge about the phylogeny of the genus in order to identify the main possible causes of phylogenetic incongruence. In **Paper I**, molecular datasets previously generated for *Medicago* are re-analysed and compared using tests based on coalescent simulation. Methods that deal with the observed incongruence, such as multi-labelling of potential hybrid taxa, are tested. Given that the molecular data produced to date is clearly insufficient to clarify species relationships in *Medicago*, the second step towards the main goal of this thesis is to develop appropriate markers that will allow for more robust phylogenetic estimations. In **Paper II**, the development of new molecular markers (low-copy nuclear loci) with the use of genomic data and novel gene-enrichment and sequencing techniques is described. These markers are evaluated to ascertain that they can be successfully captured and the respective sequences fully reconstructed, and that they contain enough information to obtain phylogenetic trees. Rates of nucleotide substitution for the proposed new markers are also inferred. As incongruence is also expected to be found among newly obtained gene phylogenies, it becomes essential to develop analytical tools that allow for a discrimination of the causes of incongruence. **Paper III** describes a methodological model that uses genomic location and coalescent simulation to discriminate between three major causes of phylogenetic incongruence: paralogy, hybridisation and incomplete lineage sorting. Once methodological improvements are available, it becomes possible to test biological questions regarding species relationships in *Medicago*. **Paper IV** addresses the question of polyploid hybridisation, revealing the existence of cryptic polyploid species in *Medicago*. **Paper V** is focused on two clades of diploid species and provides evidence of hybridisation among diploid lineages in *Medicago*, as well as a clarification of species delimitation and species relationships in the two clades.

Materials and Methods

Plant material and DNA extraction

The majority of plant material was obtained from plants grown from seed. Seeds were obtained from different sources, but the majority of accessions were obtained from the United States Department of Agriculture seeds banks (USDA/GRIN). Other major providers of seeds include Sienna Botanic Garden and SARDI (South Australia Research and Development Institute). All seeds were soaked in warm water with a small amount of detergent 24 hours before sowing. Around 10 seeds were sown in each pot, but after germination only 3-4 seedlings were kept. Plants were grown in a plant growth chamber for three to four months in order to reach frutification. Leaf material from a single individual was collected and dried with silica gel. For each collected individual, a voucher containing both vegetative and fruit material was collected and dried. In addition to plants grown from seed, leaf tissue was also obtained from herbarium specimens, namely from herbaria K, E, LD and GB. All DNA was extracted using the Dneasy Plant Mini Kit (Qiagen, Valencia, CA, USA).

Choice of markers, gene enrichment and sequencing

The choice of markers used in this thesis is described in detail in Paper II. The information available from the genome of *Medicago truncatula* (Young et al. 2011) was used to choose 61 low copy nuclear loci, each between 2 and 3 Kilo-basepairs long. Additionally, c. 200 short loci were also chosen, corresponding to single exons evenly spaced throughout all chromosomes. These shorter loci were sampled to obtain SNP data. Sequences of all the chosen markers were used to design probes intended for sequence-capture, which was carried out using the MYBaits target enrichment system (MYcroarray, Ann Arbor, Michigan). Prior to sequence capture, DNA was sheared with a Covaris S220 instrument (Covaris, Woburn, Massachusetts, USA) and DNA libraries were prepared using the NEXTflex DNA Sequencing Kit and NEXTflex Barcodes (BIOO Scientific, Austin, Texas, U.S.A). Sequencing was performed in a MiSeq platform from Illumina (San Diego, California, USA).

Data preparation and analyses

High-throughput 150 bp paired-end reads were trimmed to remove adapters and low-quality reads. Trimmed reads were then assembled or mapped onto a reference sequence. These steps were carried out using the CLC Assembly Cell software (CLC Bio, Aarhus, Denmark) available in the bioinformatics computer cluster Albiorix at the Department of Biological and Environmental Sciences, University of Gothenburg. A pipeline for data processing was generated using both bash and Python programming languages. Sequence alignments were produced using either Geneious Pro v.5.3.6 (Biomatters Ltd.) or Mafft v7.123 (Katoh and Standley 2013). Phylogenetic analyses of the data produced by next-generation sequencing are described in more detail in Papers IV and V. An additional data exploration analysis, on a concatenated set of 57 loci and 98 samples corresponding to 63 diploid taxa, was done on SplitsTree v.4 (Huson 1998), with gaps sites removed and using the default parameter (uncorrected P-distance).

Chloroplast analyses

The sequence-capture procedure utilised here, besides generating sequences of the targeted loci, worked also as a genome-skimming method which enabled the sequencing of complete or near-complete chloroplast sequences for each sample. These sequences were obtained by mapping reads

of each sample onto a reference chloroplast sequence of *Medicago truncatula* (Young et al. 2011). The alignment generated was manually corrected for misalignment and errors, and had a total size of 118 976 characters, with 83 taxa that included both *Medicago* and *Trigonella*/*Melilotus* samples. Only one individual per species was included.
A partition scheme was chosen by comparing three alternative partitionings using Bayes Factors. The three partition schemes were: two partitions (coding/non-coding), three partitions (coding/intergenic regions/regions with secondary-structure) and 78 partitions (individual coding sequences/intergenic regions/regions with secondary-structure). Models for each partition were inferred using JModelTest v. 2.1.4 (Darriba et al. 2012). A stepping-stone analysis was done for each partition scheme on MrBayes 3.2 (Ronquist et al. 2012). The Bayes Factors calculation comparing the likelihoods of each partition scheme showed no significant difference between the three partitions. The two-partition scheme was then chosen for tree inference using MrBayes. Two separate runs with two chains each were ran for 100E6 mcmc generations. The same dataset was analysed on BEAST v. 1.8 (Drummond et. al. 2012) with the chosen models (GTR + G), an uncorrelated lognormal relaxed clock prior on both partitions and a normal root height prior with mean=15.9E6 and stdev=2.7E6. The analysis ran for 100E6 mcmc generations.

Results and Discussion

Main findings

This thesis provides new evidence to explain the causes of phylogenetic incongruence in *Medicago*, demonstrating that hybridisation at the homoploid level has shaped the evolutionary history of this plant genus. It is clear that although incomplete lineage sorting (ILS) is prevalent, much of the phylogenetic incongruence encountered can be attributed to hybridisation, both homoploid and polyploid. The utility of sequence-capture techniques to solve complex phylogenetic questions is confirmed, as are coalescent-based models for sorting genomic data prior to phylogenetic analysis. The newly-generated data is used to answer questions regarding species relationships and species delimitation in two subgroups within *Medicago*.

The comparison of previously generated gene phylogenies of *Medicago* (**Paper I**) shows that these are all highly incongruent, and that the data used are insufficient to clarify all questions regarding this incongruence. Even with the reduction of those data to a common set of taxa, it was not possible to clearly discern which biological processes, other than incomplete lineage sorting, were driving the observed patterns. Nonetheless, strong hints that hybridisation was responsible for much of the incongruence observed are shown in that study. Generating the necessary data to answer this question required a completely new sampling approach. The procedure described in **Paper II** shows how the use of genomic data provided by the sequenced genome of *Medicago truncatula*, together with the recently developed sequence capture techniques, resulted in the successful sequencing of a significant number of large-sized single-copy nuclear loci. In Figure 2 a splits graph, obtained from a concatenated dataset of 57 of these loci, is presented. The splits graph shows clear signs of reticulation, which entails highly conflicting phylogenetic signal in these data. The conflict is almost certainly genome-wide and not an artefact of the initial small sample of six loci studied previously, given how many loci were sampled and that they came from every chromosome in the *M. truncatula* genome.

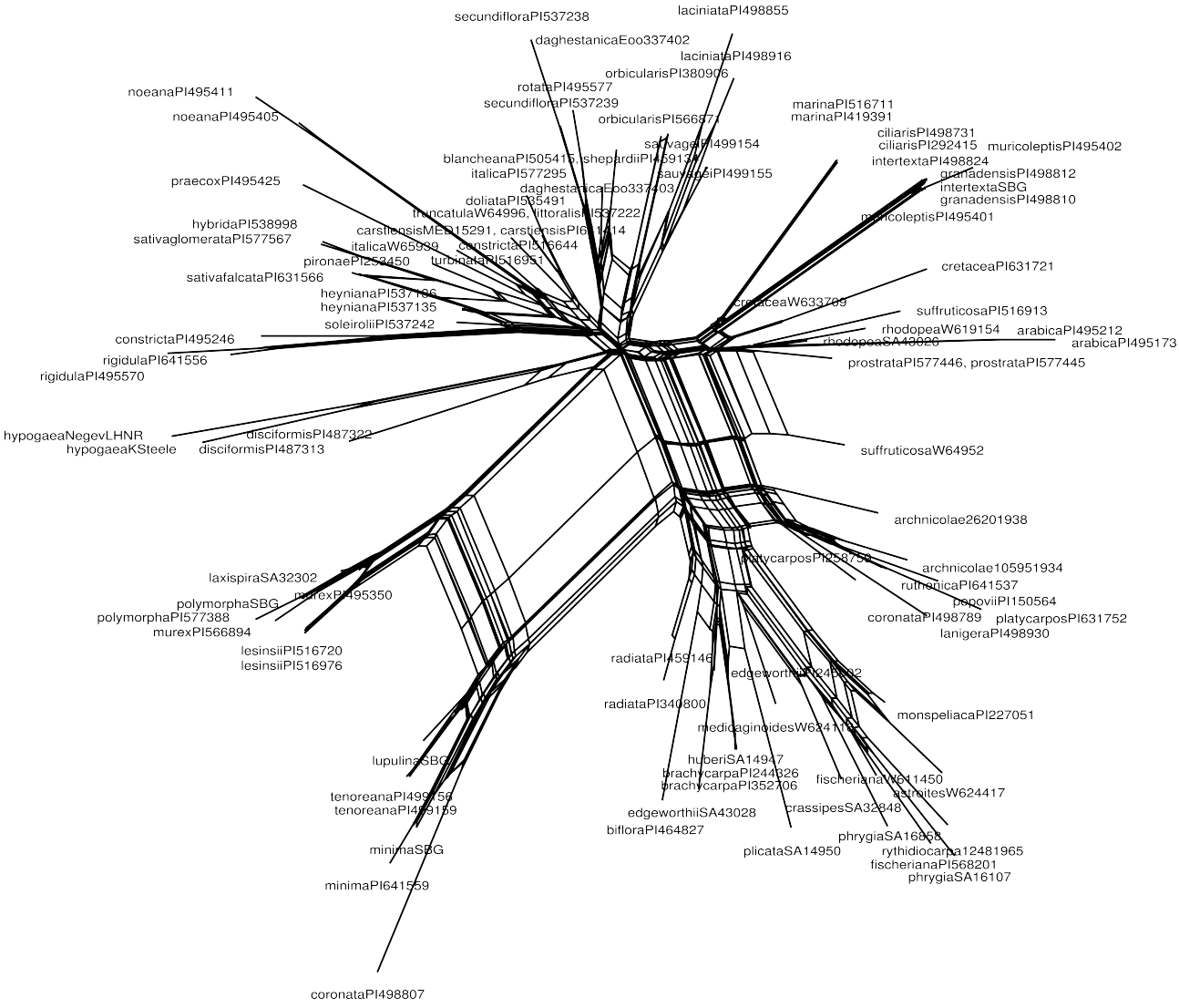

Figure 2: SplitsTree network inferred from a concatenated dataset of 57 taxa in *Medicago* using the uncorrected P distance with gap sites removed. The sampling includes 98 individuals corresponding to 63 taxa.

When using data from multiple loci and for a high number of samples, conflicting phylogenetic signal is routinely encountered (Salichos and Rokas 2013), at the very least due to the effect of incomplete lineage sorting. Detecting other sources of incongruence remains, however, a challenge. **Paper III** introduces a new model for distinguishing between incomplete lineage sorting and two other sources of phylogenetic incongruence, paralogy and (homoploid) hybridisation. It is demonstrated that using genomic location and coalescent simulation allows for the successful detection of paralogues within genomic blocks containing three low-copy nuclear genes, and for the detection of hybridisation, when genes from different genomic blocks are compared. An example of a putative ancient hybridisation event is also shown in addition to results based on simulated data. Data generated through the protocol presented in **Paper II** is used in **Paper IV** to resolve relationships within *Medicago prostrata*, revealing the presence of cryptic allopolyploidy in this species, which shows another important aspect of reticulation in *Medicago*.

The model and methods presented in **Paper III** are used in **Paper V** to investigate the phylogeny of two clades within *Medicago*. The model using genomic blocks is applied to the *M. murex* clade where two genomic blocks are identified as containing paralogous genes. The species tree and

species delimitation analyses show a clear separation between *M. murex* and *M. lesinsii*, two taxa currently considered to be the same species (Small 2011). It also suggests the possibility that two chromosomal speciation events may have occurred in the same clade and raises this question for a future investigation. Coalescent-based tree distance tests are also used successfully to sort data and obtain species trees for subsection Intertextae. Two hybridisation events are found in this clade, and the species trees obtained from the sorted data recover all four principal trees that are theoretically expected in such cases.

Utility of data and methods

Applying newly developed methods for generating data, especifically sequence-capture methods, resulted in a ten-fold increase in the amount of gene phylogenies available for *Medicago* and in the most complete sampling of the genus to date. It was also possible to obtain a fully comparable dataset, which was not the case with the dataset used in **Paper I**. The methods applied, which involve the construction of DNA libraries from a single plant extract, leave no doubt as to the identity of samples, meaning that the recovery of samples in different positions in gene phylogenies produced through these methods cannot be attributed to misidentification. Although these also occur (e.g., *M. rigidula*, **Paper V**), they play no role in the observed patterns of incongruence. As for contamination, it can never be completely ruled out, as sample preparation (e.g., DNA extraction) is usually done simultaneously on several samples, but the fact that DNA library preparation included the indexing of samples at early stages of the process, excludes the possibility of contamination downstream of that step. The effects of any possible contamination would also be corrected for after sequencing. The achieved sequencing depth and the bioinformatics procedure taken, in which sequences correspond to phased alleles, rather than consensus sequences, should allow for the separation of sequences originating in contaminating material.
The use of sequences obtained from the sequenced genome of *M. truncatula* proved to be a highly efficient way to design probes for sequence capture. These probes, which contained both intron and exon sequence, were able to successfully capture target sequences in both *Medicago* and the sister genus *Trigonella* (**Paper II**). The fact that full gene sequences were available also facilitated the processing of sequencing reads, which were compiled using a mapping approach, in which individual reads are mapped against a reference sequence. Thus, the entire data generation protocol was more efficient than similar sequence-capture protocols using exon sequences (ie., transcriptomes) as reference for probe design, which require additional steps, such as sequence assembly, in order to generate complete gene sequences. Nevertheless, building the bioinformatics pipeline used for data processing required optimising pre-existing tools that were not developed for phylogenetics. In particular, the tools used for producing contigs from individual sequence reads, most often used for analysis of single genomes, had to be modified in order to accommodate the sequence variability encountered when analysing multiple genomes.
Phylogenetic analysis of the sequenced loci resulted in resolved gene phylogenies with full support for many of the relationships, especially within sub-clades. However, the differences between these gene phylogenies are indicative of a high degree of phylogenetic conflict. Although some of the observed differences can be attributed to lack of signal, the patterns observed suggest a reticulate evolutionary history. The concatenated dataset shown in Figure 2 clearly indicates that inferring a genus-wide species phylogeny of *Medicago* from our nuclear data should not be attempted, unless violations of the coalescent, such as hybridisation, are discerned a priori. The model and tests presented in **Paper III** are based on the idea that topological distances between two gene trees can be tested for compatibility with a scenario of pure ILS. When this compatibility is rejected, other sources of incongruence, such as hybridisation and/or paralogy, are assumed to have affected the genes is question. These pairwise tests between gene phylogenies allow for the sorting of data such that conflicting signal apart from ILS is minimized, resulting in well-resolved species phylogenies.

The utility of such tests is explicitly demonstrated in **Paper IV**, where it is shown that alternative placements, in different gene phylogenies, of alleles from tetraploid taxa cannot be explained by ILS alone, supporting the hypothesis of allopolyploidy, and in **Paper V**, where a total-evidence approach, even within a coalescent framework, would not have resulted in the discovery of hybridisation in *Medicago ciliaris*. If all genes had been used in a single species tree inference analysis (for example, using *BEAST), the patterns of hybridisation identified in **Paper V** could not have been discerned, and the conflict in the data would possibly result in poorer phylogenetic estimation.

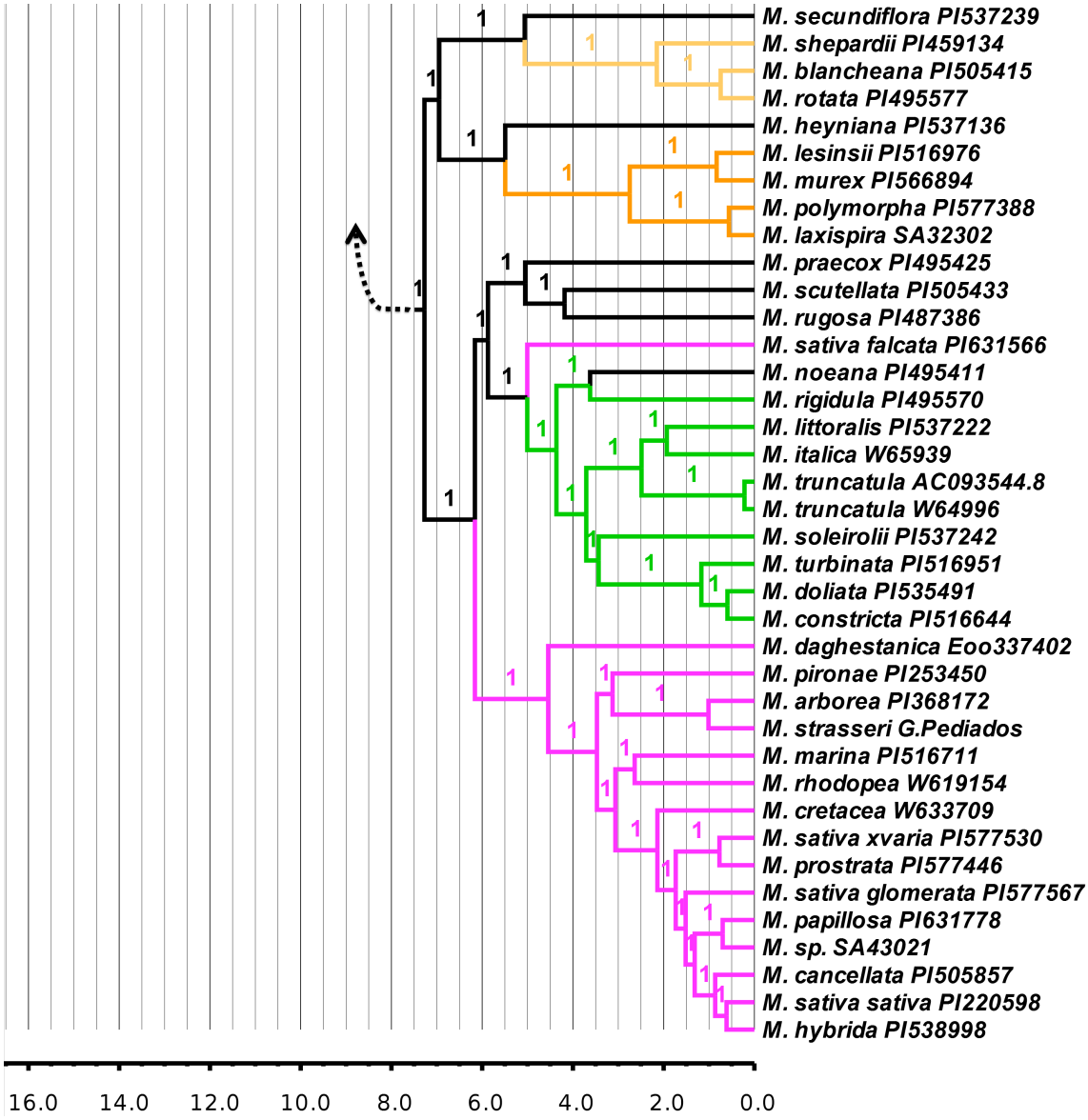

Figure 3: Phylogenetic tree inferred from a full chloroplast dataset using BEAST, with 73 species and sub-species of *Medicago* represented. The colour scheme is similar to that in Paper I.

Chloroplast data

The chloroplast analysis yielded a fully resolved and supported phylogenetic tree, with equivalent topologies obtained from the MrBayes and BEAST analyses. The tree obtained from BEAST is presented in Figure 3 and contains 75 terminals corresponding to 73 taxa. The chloroplast phylogeny is not a complete representation of the phylogenetic history of *Medicago*, which is shaped by reticulation events, but it is perhaps the only fully resolved genus-wide phylogeny that can be inferred from a single locus, or linkage group, the other alternative possibly being a mitochondrial phylogeny, which was not pursued in this thesis.

Although the relationships in this tree are fully resolved and supported, it is still not appropriate to draw conclusions on classification (e.g. monophyly of subgroups) as it still consists of a single locus. However, given that the population sizes for the chloroplast are smaller than for the nuclear genome, the timing of speciation events may be well represented in this tree. Thus, approximate

estimates of speciation rates may be inferred from the chloroplast tree, in spite of the absence of a species tree. The chloroplast tree exhibits a pattern of short branching from c. 10-6 Mya, which is suggestive of a period of accelerated speciation during that period. This aspect, however, is not explored here. *Medicago hypogaea*, a rare ephemeral from the eastern Mediterranean, appears well nested in *Medicago*, which was also observed in most nuclear gene trees. This shows that this species, once in its own genus (*Faktorovskya* Eig), is indeed part of *Medicago*.

Hybridisation in Medicago

Natural hybridisation between diploid species of *Medicago* is demonstrated in **Paper V**, confirming the hypotheses elaborated in **Paper I** and in earlier papers on the phylogeny of *Medicago* (e.g., Maureira-Butler et al. 2008). Species delimitation analyses in **Paper V** indicate, without doubt, that the individual referred to as *Medicago ciliaris* 731 has received genetic contributions from the lineage containing *M. intertexta* and *M. muricoleptis*. The pattern is consistent with a scenario of introgression from this lineage into *M. ciliaris*, as this relationship appears recent. The sample referred to as *Medicago ciliaris* 415, on the other hand, appears to be a hybrid between the *M. ciliaris* lineage and an unsampled or extinct lineage that is sister to the entire subsection Intertextae. This sample may or not correspond to a case of homoploid hybrid speciation, but it is certain that it received contributions from two distinct lineages. Reinforcing the idea of two hybridisation events in this clade is the fact that species tree inference recovered the four principal trees expected in such cases, each principal representing a different combination of relationships for each of the hybrid taxa. Other cases of introgression, involving other species, may also be present in this clade, albeit in a minor scale. Ancient hybridisation between diploid lineages of *Medicago* was also suggested in **Paper III.** According to the coalescent-based model and tests described in this paper, the position of *M. orbicularis* in relation to two distinct lineages of *Medicago* cannot be explained by a scenario of ILS alone. However, this hypothesis needs to be tested further with the use of more genes.
Using the markers presented in **Paper II** it was also possible to show that allopolyploidy, another form of hybridisation, has originated several lineages in *Medicago*, including cryptic allopolyploid species (**Paper IV**; Eriksson et al. in prep.). In Paper IV it is demonstrated that individuals of *Medicago prostrata* with tetraploid chromosome counts have an allopolyploid origin, derived from hybridisation between the *M. prostrata* lineage and the *M. sativa* lineage. It is clear that species of *Medicago* with two chromosome counts (diploid and polyploid) can in fact include allopolyploid lineages, which shows the complexity of reticulation in *Medicago*.

Future prospects

Many questions remain regarding the evolutionary history of *Medicago*. The importance of hybridisation in shaping this history is now demonstrated, and relationships within certain groups have been clarified, but several subgroups within the genus remain to be studied. For example, relationships in the *Medicago truncatula* and *M. sativa* clades need further investigation, as gene phylogenies suggest the possibility that hybridisation has also affected these clades. Research on hybridisation within *Medicago sativa* itself is also of the utmost importance, given the global importance of that species complex, and will possibly require studies at the population level. The origin of certain polyploid lineages, such as the *Medicago rugosa*/*scutellata* lineage, has not been fully understood. In order to make full use of sequence data for the study of polyploids, new algorithms, allowing for the automatic separation of all alleles present in polyploid individuals, must be developed. Adjustments to the currently used protocols may be necessary in order to obtain the optimal read depth that would enable the recovery of all these alleles, depending on the level of polyploidy.

The use of genomic blocks for understanding hybridisation among diploids can be further tested, preferably with a more ambitious approach than the one followed in this thesis. Namely, targeting larger genomic blocks, in higher number and for a higher number of individuals from the same species and populations will allow for a better understanding of the dynamics of hybrid formation and the role of adaptation for the fixation of genomic material in hybrid populations.
The chloroplast dataset has not been fully explored. As all relationships in the chloroplast phylogeny are resolved and supported, these data may be the most adequate to infer the past dynamics of speciation in *Medicago.* For example, if different speciation rates are found, this may indicate changes in past environmental conditions driving speciation and extinction in the genus. One relevant question to be addressed using this information would be whether *Medicago* has been affected by geological events such as the Messinian salinity crisis, which occurred 5-6 My ago, after the estimated split between *Trigonella* and *Medicago*. Testing for the possibility of a rapid radiation having occurred early in the *Medicago* lineage may also help to clarify the incongruent patterns observed deep in the tree, between the major lineages, and whether or not hybridisation has also played a role in shaping those patterns of incongruence.

Acknowledgements

A very enthusiastic, meritorious and really imperative appreciation
of all the support that made this thesis possible,

I thank:

Bernard Pfeil, my supervisor, who is always very optimistic, encouraging and patient,

Bengt Oxelman, who always gives me precious advice,

Yann Bertrand, who accepts all challenges with loyalty and dedication,

Mari Källersjö, my examiner, for her support and sympathy,

Elisabet Sjökvist, Anna Petri and Stephan Nylinder, who were very welcoming and helpful when I came to Gothenburg,

Zeynep Aydin, for her unshakable kindness,

Bente Eriksen and Claes Persson, for all the botanical knowledge I gained from them,

Vivian Aldén and Sven Toresson, who make everything work at Botan,

Jeff Doyle, for his kind contribution to Paper III,

Mats Töpel, who is always alert and ready to solve problems with the cluster,

Henrik Nilsson, who is a true source of motivation,

and all those who contributed directly or indirectly to the completion of this thesis:

Thomas H., Karine B., Alexander Z., Ivana K., Daniela A., Ylva H., Gustavo H., Jonna E., Alexandre A., Roger E., Ellen L., Marcela F., Rosemeri M., Graça, Lennart and Catarina P., Tobias H., José B-P., Isabel L., Sazzad K., Fabian R., Cajsa A., Thomas M., …

References

Bauchan, G., & Greene, S. (2000). Report on the Status of *Medicago* Germplasm in the United States. Alfalfa Crop Germplasm Committee. In *37th North American Alfalfa Improvement Conference Madison, WI, July* (pp. 16-19).

Bena, G. (2001). Molecular phylogeny supports the morphologically based taxonomic transfer of the "Medicagoid" *Trigonella* species to the genus *Medicago* L.. Plant Systematics and Evolution, 229, 217-236.

Bena, G., Jubier, M., Olivieri, I. & Lejeune, B. (1998). Ribosomal External and Internal Transcribed Spacers: Combined Use in the Phylogenetic Analysis of *Medicago* (Leguminosae). Journal of Molecular Evolution, 46, 299-306.

Bertrand, Y. J., Scheen, A. C., Marcussen, T., Pfeil, B. E., de Sousa, F., & Oxelman, B. (2015). Assignment of Homoeologues to Parental Genomes in Allopolyploids for Species Tree Inference, with an Example from *Fumaria* (Papaveraceae). *Systematic Biology*, syv004.

Blanc, G. & Wolfe, K. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. The Plant Cell, 16, 1667-1678.

Bock, D. G., Kane, N. C., Ebert, D. P. & Rieseberg, L.H. (2014). Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. New Phytologist, 201, 1021-1030.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A. & RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Molecular Biology and Evolution, 29, 1917-1932.

Carstens, B. C., Brennan, R. S., Chua, V., Duffie, C. V., Harvey, M. G., Koch, R. A., McMahan, C. D., Nelson, B. J., Newman, C. E., Satler, J. D., Seeholzer, G., Posbic, K., Tank, D. C. & Sullivan, J. (2013). Model selection as a tool for phylogeographic inference: an example from the willow *Salix melanopsis*. Molecular Ecology, 22, 4014-4028.

Cronn, R., Knaus, B. J., Liston, A., Maughan, P. J., Parks, M., Syring, J. V. & Udall, J. (2012). Targeted enrichment strategies for next-generation plant biology. American Journal of Botany, 99, 291-311.

Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J., Soltis, P. S., Carlson, J. E., Arumuganathan, K., Barakat, A., Albert, V. A., Ma, H. & dePamphilis, C.W. (2006). Widespread genome duplications throughout the history of flowering plants. Genome Research, 16, 738-749.

Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. Nature Methods, 9, 772-772.

Degnan, J. H. & Rosenberg, N.A. (2006). Discordance of species trees with their most likely gene trees. PLoS Genetics, 2, e68.

Downie, S. R., Katz-Downie, D. S., Rogers, E. J., Zujewski, H. L. & Small, E. (1998). Multiple

independent losses of the plastid rpoC1 intron in *Medicago* (Fabaceae) as inferred from phylogenetic analyses of nuclear ribosomal DNA internal transcribed spacer sequences. Canadian Journal of Botany, 76, 791-803.

Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. (2012). Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution, 29, 1969-1973.

Eaton, D. A. R. & Ree, R.H. (2013). Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (*Pedicularis*: Orobanchaceae). Systematic Biology, 62, 689-706.
Gaston, K. J. & Spicer, J.I. (2013). Biodiversity: an introduction. John Wiley \& Sons.

Gur, A. & Zamir, D. (2004). Unused natural variation can lift yield barriers in plant breeding. PLoS Biology, 2, e245.

Hajjar, R. & Hodgkin, T. (2007). The use of wild relatives in crop improvement: a survey of developments over the last 20 years. Euphytica; Netherlands Journal of Plant Breeding, 156, 1-13.
Holland, B., Benthin, S., Lockhart, P., Moulton, V. & Huber, K. (2008). Using supernetworks to distinguish hybridization from lineage-sorting. BMC Evolutionary Biology, 8, 202.

Hughes, C. E., Eastwood, R. J. & Bailey, C.D. (2006). From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. Philosophical Transactions of the Royal Society B: Biological Sciences, 361, 211-225.

Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data.. Bioinformatics (Oxford, England), 14, 68-73.

Huyghe, C., Bertin, E., Landry, N., Acharya, S., Thomas, J. & others (2007). Medicinal and nutraceutical uses of alfalfa (*Medicago sativa* L.).. Advances in medicinal plant research 2007., , 147-172.

Ivanov, A. I. (1977). History, origin and evolution of the genus *Medicago* of the subgenus Falcago. *Trudy po prikladnoĭ botanike, genetike i selekt͡sii͡*=.

Joly, S., McLenachan, P. A. & Lockhart, P.J. (2009). A Statistical Approach for Distinguishing Hybridization and Incomplete Lineage Sorting.. The American Naturalist, 174, pp. E54-E70.

Jones, G., Sagitov, S. & Oxelman, B. (2013). Statistical Inference of Allopolyploid Species Networks in the Presence of Incomplete Lineage Sorting. Systematic Biology, 62, 467-478.

Katoh, K. & Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution, 30, 772-780.

Kingman, J. (1982). The coalescent. Stochastic Processes and their Applications, 13, 235 – 248.

Kubatko, L. S. (2009). Identifying Hybridization Events in the Presence of Coalescence via Model Selection. Systematic Biology, 58, 478-488.

Kubatko, L. S. & Degnan, J.H. (2007). Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. Systematic Biology, 56, 17-24.

Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefings in Bioinformatics, 5, 150-163.

Lemmon, A. R. & Lemmon, E.M. (2012). High-Throughput Identification of Informative Nuclear Loci for Shallow-Scale Phylogenetics and Phylogeography. Systematic Biology, 61, 745-761.

Lemmon, A. R., Emme, S. A. & Lemmon, E.M. (2012). Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Systematic Biology, 61, 727-744.

Lewis, G. P., Schrire, B., Mackinder, B. & Lock, M. (2005). Legumes of the World. Royal Botanic Gardens Kew.

Liu, L. & Pearl, D.K. (2007). Species Trees from Gene Trees: Reconstructing Bayesian Posterior Distributions of a Species Phylogeny Using Estimated Gene Tree Distributions. Systematic Biology, 56, 504-514.

Lynch, M. & Conery, J.S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. Science, 290, 1151-1155.

Mal\'e, P. G., Bardon, L., Besnard, G., Coissac, E., Delsuc, F., Engel, J., Lhuillier, E., Scotti-Saintagne, C., Tinaut, A. & Chave, J. (2014). Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. Molecular ecology resources, 14, 966-975.
Maureira-Butler, I. J., Pfeil, B. E., Muangprom, A., Osborn, T. C. & Doyle, J.J. (2008). The Reticulate History of *Medicago* (Fabaceae). Systematic Biology, 57, 466-482.

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C. & Brumfield, R.T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution, 66, 526 – 538.

Meng, C. & Kubatko, L.S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. Theoretical Population Biology, 75, 35 – 45.

Mikaili, P., & Shayegh, J. (2011). *Medicago sativa*: a historical ethnopharmacology and etymological study of the alfalfa.

Oberprieler, C. (2005). Temporal and spatial diversification of circum-Mediterranean Compositae-Anthemideae. Taxon, 54, 951-966.

Ogden, T. H. & Rosenberg, M.S. (2006). Multiple sequence alignment accuracy and phylogenetic inference. Systematic Biology, 55, 314-328.

Petri, A. & Oxelman, B. (2011). Phylogenetic relationships within *Silene* (Saryophyllaceae) section Physolychnis. Taxon, , 953-968.

Quiros, C. F. (1988). 3 The Genus *Medicago* and the Origin of the *Medicago sativa* ComplexI.

Rannala, B. & Yang, Z. (2003). Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci. Genetics, 164, 1645-1656.

Rasmussen, M. D. & Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence

using a locus tree. Genome Research, 22, 755-765.

Rieseberg, L. H. (1991). Homoploid reticulate evolution in *Helianthus* (Asteraceae): evidence from ribosomal genes. American Journal of Botany, , 1218-1237.

Rieseberg, L. H. (1997). Hybrid Origins of Plant Species. Annual Review of Ecology and Systematics, 28, pp. 359-389.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. & Huelsenbeck, J.P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology, 61, 539-542.

Salichos, L. & Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. Nature, 497, 327-331.

Sang, T., & Zhong, Y. (2000). Testing hybridization hypotheses based on incongruent gene trees. Systematic Biology, 49(3), 422-434.

Small, E. (1996). Adaptations to herbivory in alfalfa (*Medicago sativa*). Canadian journal of botany, 74, 807-822.

Small, E. (2011). Alfalfa and relatives. Evolution and classification of *Medicago*. NRC Research Press, Ottawa.

Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C., & Brumfield, R. T. (2013). Target capture and massively parallel sequencing of ultraconserved elements (UCEs) for comparative studies at shallow evolutionary time scales.*Systematic biology*, syt061.

Steele, K. P., Ickert-Bond, S. M., Zarre, S. & Wojciechowski, M.F. (2010). Phylogeny and character evolution in *Medicago* (Leguminosae): Evidence from analyses of plastid trnK/matK and nuclear GA3ox1 sequences. American Journal of Botany, 97, 1142-1155.

Stull, G. W., Moore, M. J., Mandala, V. S., Douglas, N. A., Kates, H., Qi, X., Brockington, S. F., Soltis, P. S., Soltis, D. E. & Gitzendanner, M.A. (2013). A Targeted Enrichment Strategy for Massively Parallel Sequencing of Angiosperm Plastid Genomes. Applications in Plant Sciences, 1, 1-7.

Swofford, D. L., Olsen, G. J., Waddell, P. J., & Hillis, D. M. (1996). {Phylogenetic inference}.

Wang, X. Q., Zhao, L., Eaton, D. A. R., Li, D. Z. & Guo, Z.H. (2013). Identification of SNP markers for inferring phylogeny in temperate bamboos (Poaceae: Bambusoideae) using RAD sequencing. Molecular Ecology Resources, 13, 938-945.

Wendel, J. F. (2000). Genome evolution in polyploids. In *Plant molecular evolution* (pp. 225-249). Springer Netherlands.

Wiens, J. J. (2007). Species delimitation: new approaches for discovering diversity. Systematic Biology, 56, 875-878.

Yoder, J. B., Briskine, R., Mudge, J., Farmer, A., Paape, T., Steele, K., Weiblen, G. D., Bharti, A. K.,

Zhou, P., May, G. D., Young, N. D. & Tiffin, P. (2013). Phylogenetic Signal Variation in the Genomes of *Medicago* (Fabaceae). Systematic Biology, 62, 424-438.
Young, N. D., Debelle, F., Oldroyd, G. E. D., Geurts, R., Cannon, S. B., Udvardi, M. K., Benedito, V. A., Mayer, K. F. X., Gouzy, J., Schoof, H., Van de Peer, Y., Proost, S., Cook, D. R., Meyers, B. C., Spannagl, M., Cheung, F., De Mita, S., Krishnakumar, V., Gundlach, H., Zhou, S., Mudge, J., Bharti, A. K., Murray, J. D., Naoumkina, M. A., Rosen, B., Silverstein, K. A. T., Tang, H., Rombauts, S., Zhao, P. X., Zhou, P., Barbe, V., Bardou, P., Bechner, M., Bellec, A., Berger, A., Berges, H., Bidwell, S., Bisseling, T., Choisne, N., Couloux, A., Denny, R., Deshpande, S., Dai, X., Doyle, J. J., Dudez, A., Farmer, A. D., Fouteau, S., Franken, C., Gibelin, C., Gish, J., Goldstein, S., Gonzalez, A. J., Green, P. J., Hallab, A., Hartog, M., Hua, A., Humphray, S. J., Jeong, D., Jing, Y., Jocker, A., Kenton, S. M., Kim, D., Klee, K., Lai, H., Lang, C., Lin, S., Macmil, S. L., Magdelenat, G., Matthews, L., McCorrison, J., Monaghan, E. L., Mun, J., Najar, F. Z., Nicholson, C., Noirot, C., O/'Bleness, M., Paule, C. R., Poulain, J., Prion, F., Qin, B., Qu, C., Retzel, E. F., Riddle, C., Sallet, E., Samain, S., Samson, N., Sanders, I., Saurat, O., Scarpelli, C., Schiex, T., Segurens, B., Severin, A. J., Sherrier, D. J., Shi, R., Sims, S., Singer, S. R., Sinharoy, S., Sterck, L., Viollet, A., Wang, B., Wang, K., Wang, M., Wang, X., Warfsmann, J., Weissenbach, J., White, D. D., White, J. D., Wiley, G. B., Wincker, P., Xing, Y., Yang, L., Yao, Z., Ying, F., Zhai, J., Zhou, L., Zuber, A., Denarie, J., Dixon, R. A., May, G. D., Schwartz, D. C., Rogers, J., Quetier, F., Town, C. D. & Roe, B.A. (2011). The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature, 480, 520-524.

Şakiroğlu, M., Doyle, J. J. & Brummer, E.C. (2010). Inferring population structure and genetic diversity of broad range of wild diploid alfalfa (*Medicago* sativa L.) accessions using SSR markers. Theoretical and applied genetics, 121, 403-415.