Polyploid Phylogenetics in Plants: Insights on non-model organisms in Fabaceae and Malvaceae

Jonna Sofia Eriksson

2019

UNIVERSITY OF GOTHENBURG

Faculty of Science

Department of Biological and Environmental Sciences

Opponent: Dr. Jonathan F. Wendel

Examiner: Prof. Mari Källersjö

Supervisors: Prof. Alexandre Antonelli, Prof. Bengt Oxelman

Eriksson, J. S. (2019): Polyploid Phylogenetics in Plants: Insights on non-model organisms in Fabaceae and Malvaceae. PhD thesis. Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden.

Cover image: © Jonna S. Eriksson

Copyright ISBN print: 978-91-7833-241-0
ISBN digital: 978-91-7833-242-7
Digital version available at http://hdl.handle.net/2077/59804
Printed by BrandFactory AB

“All generalizations are dangerous,

even this one”

- Alexandre Dumas

Table of Contents

Abstract 6

Svensk sammanfattning 10

Publications included 14

Introduction to polyploidy 16

Objectives 26

Materials and Methods 28

Study groups 28

DNA extraction and sequencing 33

Results and Discussion 36

Conclusions 44

Future prospects 46

Paper contributions 48

References 50

Acknowledgements 58

Abstract

Genome duplication is a common phenomenon in angiosperms and advances in sequencing technologies and bioinformatics is now revealing its prevalence and significance. In a polyploidization event all genes in a genome are doubled. Duplicated genes can take one of three paths, either both of the duplicated are maintained or one of them may be randomly lost or selected against. Polyploid species are challenging for resolving species relationships because of the number of duplicated genes and the different processes leading to genome reduction.

In this thesis I investigate the mode of polyploid origin (i.e. auto- and allopolyploidy) and the role of ancient genome duplication in plant speciation. Here, two families were studied because of their numerous rounds of polyploidization events, Fabaceae and Malvaceae. To discern between the complex processes involving polyploidy, large amounts of data were generated from several nuclear genes using target gene capture and Illumina sequencing.

In *Medicago* (Fabaceae) two modes of polyploidy (autopolyploidy and allopolyploidy) were discovered. In the first case, an autopolyploid mode was identified, because gene comparisons showed two independently evolving species, a cryptic tetraploid species derived from a diploid progenitor. In the second case, two woody tetraploids were found to be hybrids. Since the closest parental lineages associated with the hybridization only have woody roots, the genome duplication for genes related to woodiness may be an instance of transgressive phenotype (extreme morphological character) in the hybrids.

Evidence of genome duplication appears not only in recently formed polyploids, but may also be present through genes that were duplicated in a lineage's history. The traces of ancient genome duplication events may be

scarce owing to random mutations and gene losses. *Hibiscus* (Malvaceae) is a plant genus which possesses diverse chromosome numbers among species, indicative of potential polyploidization events. By studying the number of gene copies in diploid species, two ancient genome duplications were identified in *Hibiscus*. Additionally, numerous polyploidization events following the ancient duplications were detected among the extant species, indicating a complex reticulate history.

Hibisceae consists of five major clades: /Calyphylli, /Euhibiscus, /Furcaria, /Trionum and /Megistohibiscus but with inconsistent genus naming conventions (e.g. *Hibiscus* occurs across all of the clades). In this study, phylogenetic analysis of HTS data supported the classification of the group into the major clades; all were found to be monophyletic and no hybrid polyploidization events were found to have occurred between them. Additionally, each major clade's taxa were found to have common base chromosome numbers. Given the results, a taxonomic renaming, based on base chromosome number, of the major clade's genera is recommended.

This thesis demonstrates a rich history of polyploidizations both recent and ancient – highlighting the important role this phenomenon has played in the evolution of two distantly-related plant families. Polyploidy may explain the underlying causes of when classical taxonomy (classification before DNA sequences) is not enough and may potentially lead to underestimation of the true number of species. Due to the unique patterns across lineages, polyploidizations allows for no generalizations; despite its ubiquity, it remains mysterious. Polyploidy is, at least in part, reversible and leads to a smaller genome size over time.

Keywords: Diploidization, genome duplication, haplotypes, Hibisceae, *Hibiscus*, homoeologues, high-throughput sequencing, *Pavonia*, Malvaceae, polyploidization

Svensk sammanfattning

Genomduplicering, när antalet kromosomer fördubblas, är ett vanligt förekommande fenomen hos blomväxter. Man tror att 15% av alla blomväxter har genomgått genomdupliceringar, också kallat polyploidisering, och att 31% av alla arter inom ormbunkar är polyploida. Även om fenomenet är mindre vanligt hos djur så har man exempelvis funnit vissa belägg för att det förekommit långt bak i tiden hos ryggradsdjur. Polyploidisering kan ske inom en art och kallas då autopolyploidi, eller mellan olika arter via hybridisering, känt som allopolyploidi. Ett sätt att undersöka polyploidi hos växter är att räkna antalet kromosomer. Polyploidi resulterar i exakta dubbleringar av kromosomer. Dock kan man inte utröna vilken process (d.v.s. auto- eller allopolyploidi) som ligger till grund för fenomenet utifrån kromosomantalet. Istället används genetiska data, såsom DNA, för att rekonstruera arternas släktskap. Detta kan man göra genom att jämföra DNA:t mellan arter och konstruera ett dikotomt förgrenat släktskapsträd, där arter som är närbesläktade förekommer tillsammans. En art som är resultatet av autopolyploidi kommer att uppträda som systergrenar i trädet, medan en art med ett allopolyploid ursprung kommer att grupperas tillsammans med respektive stamfader.

Utöver genomdupliceringar, så finns det en process som reducerar antalet kromosomer och deras gener. Diploidisering gör att arter med flera uppsättningar av kromosomer (t.e.x. tetraploida har fyra uppsättningar av kromosomer) kommer att reduceras över tid och likna sina diploida släktingar (två uppsättningar kromosomer). Den reducerande processen visar sig vara lika förekommande som antalet dupliceringar, men är oftast svårare att urskilja. Det vill säga, en reducering av antalet kromosomer och gener gör att informationen försvinner och blir därmed oåtkomlig för DNA-jämförelser. Samtidigt resulterar det i att arter som har haft ett polyploid ursprung kan ha

kromosomantal som tolkas som om de var diploida. Problemet visar sig tydligast när man studerar artbildningen med DNA och rekonstruerar ett släktskapsträd, där diploida och polyploida arter kan ha flera varierande placeringar p.g.a. deras uppkomst (auto- eller allopolyploidy) eller för att gener har försvunnit. Det sistnämnda är ett stort problem när man vill jämföra gener som har samma ursprung. För att komma till rätta med problematiken kan man undersöka många gener och leta efter skillnader i arternas placering i trädet.

I avhandlingen undersöks de processer som är associerade med polyploidi, dess formation och hur forntida genomdupliceringar komplicerar bilden av nutida arters uppkomst. Två växtfamiljer, Fabaceae (ärtfamiljen) och Malvaceae (Malvafamiljen) var utvalda p.g.a. de numerära dupliceringar som har förekommit i båda familjerna, vilket resulterat i flertalet polyploidi-arter. I studie 1 upptäcktes båda formerna av polyploidi, auto- och allopolyploidi inom Fabaceae. Den förstnämnda visade sig vara en ny kryptisk art, *Medicago tetraprostrata*, som dolde sig tillsammans med sin diploida anfader, *M. prostrata*. Inga synliga morfologiska kännemärken kunde hittas mellan arterna, men det dubbla kromosomantalet (tetraploid jämfört med de diploida individerna) tillsammans med dess genetiska komposition visade sig vara separat och självständigt utvecklat från anfader.

I studie 2 undersökte vi ursprunget för två tetraploida växtarter som också har en distinkt morfologi. *Medicago arborea* och *M. strasseri* är de enda förekommande ved-liknande buskarna i släktet *Medicago*, där majoriteten av arterna är örtartade. Deras ursprung är intressant i ett evolutionärt perspektiv. Är deras ved-liknande struktur nedärvd från en okänd anfader och har resten av *Medicago* förlorat egenskapen? Eller har anfadern varit örtlik och den ved-liknande strukturen uppkommit individuellt i buskarna? Det visade sig att anfadern inte var vedaktig utan att de två tetraploida *M. arborea* och *M. strasseri* erhöll egenskapen genom en hybridisering som resulterade i

genomduplicering. Vilka som är de direkta föräldrarna kunde inte avgöras i släktskapsträdet, förutom att de närmsta systerarterna har egenskapen även om den inte uttryckts i samma omfattning som hos hybriderna.

Genomdupliceringar förekommer inte bara i nu levande arter, utan kan även ha skett längre bak i tiden. Genom slumpvisa mutationer och bortfall av gener så försvinner delar av den information som visar på forntida polyploidi-händelser. Detta upptäcktes i studie 3 inom växtfamiljen Malvaceae. Även om antalet kromosomer indikerade möjliga förhistoriska dupliceringar och senare reduktioner, kunde två dupliceringar fastställas i *Hibiscus* (Malvaceae) historia genom att undersöka de numerära förekomsterna av duplicerade gener. Därefter har flera artbildningar skett genom individuella polyploidi-tillfällen som kan ligga till grund för varför klassificeringen av *Hibiscus* visar sig vara otydlig mellan olika gener och dess morfologiska karaktärer. Detta kunde urskiljas i studie 4, som jämförde flera gener och arter inom *Hibiscus* för att förstå bakgrunden till varför föregående studier inte har varit framgångsrika.

Genomdupliceringar ger förutsättning för ny artbildning inom blomväxter. Det som blivit alltmer tydligt genom studierna är att arter med polyploid natur ofta gör saker lite annorlunda, d.v.s. att det inte går att generalisera mellan arter, släkten eller familjer. Processerna involverade i genomdupliceringar kommer för det mesta skilja sig åt hos olika växter och vetskapen om detta kommer att sysselsätta även framtida systematiker som arbetar med blomväxter.

Publications included

This thesis is based on the following papers, referred to in the text by Roman numerals. Paper I is reprinted with permission from Elsevier and Paper II under CC-BY license.

I. **J. S. Eriksson**, J. L. Blanco-Pastor, F. Sousa, Y. J. K. Bertrand, B. E. Pfeil (2017). A cryptic species produced by autopolyploidy and subsequent introgression involving *Medicago prostrata* (Fabaceae). Molecular Phylogenetics and Evolution 107: 367-381. https://doi.org/10.1016/j.ympev.2016.11.020

II. **J. S. Eriksson**, F. Sousa, Y. J. K. Bertrand, A. Antonelli, B. Oxelman, B.E. Pfeil (2018). Allele phasing is critical to revealing a shared allopolyploid origin of *Medicago arborea* and *M. strasseri* (Fabaceae). *BMC Evolutionary Biology* 18:9. DOI 10.1186/s12862-018-1127-z

III. **J. S. Eriksson**, D. J. Bennett, C. D. Bacon, B. E. Pfeil, B. Oxelman, A. Antonelli (submitted to GBE). Two ancient genome duplication events shape diversity in *Hibiscus* L. (Malvaceae).

IV. **J. S. Eriksson**, V.M. Gonçalez, D. J. Bennett, C. D. Bacon, B. E. Pfeil, B. Oxelman, A. Antonelli (manuscript). Base chromosome number variations and major polyploidization events impact taxonomic classification in Hibisceae.

Introduction to polyploidy

Flowering plants are fundamentally polyploid in origin, meaning that they contain more than two sets of chromosomes in their nucleus. The accelerated rate of discovery of polyploids in angiosperms and in eukaryotes has led to a greater understanding of this phenomenon and its importance across the tree of life – as it often leads to instantaneous speciation (Orr, 1990, Martino and Sinsch, 2002). To date, polyploidy is recognized as the most significant force of speciation in plants (Stebbins, 1947, Stebbins, 1969, Soltis and Soltis, 1995, Otto and Whitton, 2000, Jiao et al., 2012, Kim et al., 2017) involving some of the most important agricultural crops (Renny-Byfield and Wendel, 2014).

The definition of polyploidy – There are two primary modes of polyploid formation, *autopolyploidy* and *allopolyploidy*, and both incur genome doubling (Stebbins Jr, 1950, Grant, 1981). These modes can be defined in two ways:

> The *taxonomic* definition, where one or two parental species contribute genomic information to the polyploid daughter (Soltis et al., 2010). Autopolyploidy entails genome doubling within a single species and contains the identical replicated genome as the parent. Allopolyploidy, on the other hand, is formed by genome duplication through hybridization between two species, thus, they have two diverged genomes inherited from the progenitors. The intermediate state is referred to as *segmental allopolyploidy* and occurs when two relatively similar but distinct species hybridize (Stebbins, 1947). A segmental polyploid will have a genome that in parts will resemble autopolyploids due to similar regions, whereas other regions of the genome are so different that it resembles an allopolyploid.

In the *genetic* definition, chromosome pairing is used to define the polyploid origin. When a genome is duplicated through autopolyploidy, the chromosomes form multivalent pairs (more than one homologous chromosome pair along their length), allowing recombination to exchange genetic material between any pair of chromosomes. Recombination occurs when homologous chromosomes pair. The homologues are identical in autopolyploids and have equal opportunities to pair (Wu et al., 2001), which may result in a quadrivalent configuration, or they form multivalent pairs with different sets every generation (Havananda et al., 2011). For allopolyploids originating from two different genomes, the set of chromosomes will not pair up randomly, but instead form bivalents (only homologous chromosomes can pair) with respective parental genome (Roux and Pannell, 2015). Recombination, therefore, will exchange genetic material within each parental set of chromosomes – within each unique genome – rather than between the sets. The intermediate state, segmental allopolyploidy, displays instead both multivalent and bivalent chromosome pairing due to their partly similar and dissimilar genome regions. Distinguishing between the two definitions are not easy.

Polyploidy has commonly been inferred by comparing chromosome numbers among closely related species as they often show a doubled pattern (e.g. ploidy level ranging from $2n = 12, 24, 48$; Soltis et al., 2010). This has mainly been done using chromosome counting under light microscope or with flow cytometry. However, obtaining the information of ploidy level or chromosome number with these techniques may be insufficient for the determination of their mode of origin (Otto and Whitton, 2000). Phylogenetic studies, on the other hand, can infer the mode of polyploid origin either via autopolyploidy or

allopolyploidy, but cannot give any information on the exact number of chromosomes. Genes that are duplicated through polyploidy are referred to as homoeologues (allopolyploidy) or ohnologues (autopolyploidy; in honor of the late Susumu Ohno; Ohno, 1970, Wolfe, 2001). In a phylogenetic tree, ohnologues are expected to group together at any given locus (Fig. 1a), whereas the homoeologues are expected to be sister to the parental lineage it originated from (Fig. 1b). In this thesis, a group of alleles/genes that occur on the same chromosome inherited from a single parent and without known origin are referred to as haplotypes.

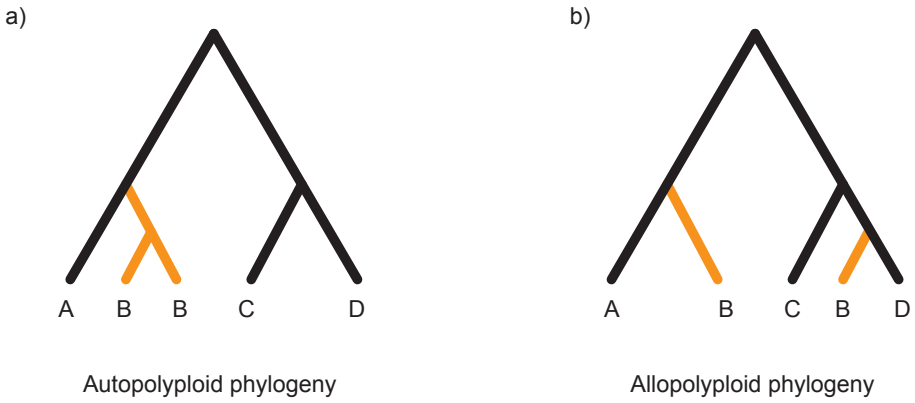


Figure 1. Phylogenetic trees that demonstrate the expected placement of haplotypes from polyploid formation. In a) ohnologous copies (shown as B tip labels), are branching next to each other, whereas b) the two homoeologous B-copies are sister to either parental lineages.

Ancient genome duplication and diploidization – Detecting the mode of polyploid origin is challenging, but identifying ancient genome duplications is even harder. Recent polyploids (neopolyploidy) may be identified by their chromosome number, genome size, intermediate morphology and gene copy number (Sémon and Wolfe, 2007, Salmon and Aïnoiche, 2015). Ancient (paleopolyploidy) genome duplication events, however, lose these signals of

duplication over time (Sémon and Wolfe, 2007). Gene copies formed by polyploidization events may also only have a brief lifespan before one copy is lost (Sémon and Wolfe, 2007) – leaving the other copy in a single-copy state. The "paradox" of returning a duplicated genome to a diploid state is referred to as *diploidization* (Soltis et al., 1993), and involves gene loss, genomic rearrangements and genome downsizing (Sémon and Wolfe, 2007, Wendel, 2015). When an established polyploid genome transitions back toward a diploid-like state (Blanc and Wolfe, 2004, Soltis et al., 2015, Pellicer et al., 2018), it causes the initially identical or nearly identical DNA from the whole-genome duplication event to diverge by random mutation and/or selection (Sankoff and Zheng, 2018). Duplicated copies may take on new functions or copies may lose part of their function and only function as a pair, where a loss of either copy is lethal (Doyle et al., 2008). Other duplicated genes may have been lost completely (Sankoff and Zheng 2018).

The probability for both duplicated genes to become fixed may be unequal following a polyploidization event (Sémon and Wolfe, 2007). The loss of copies affects all genes, but one parental genome may be preferentially retained. This phenomenon is referred to as *fractionation*, which can either occur at random or through selection (Soltis et al., 2015, Wendel, 2015). This form of selective loss obscures one genomic parent (Doyle et al., 2008), biasing our understanding of genome evolution and causing incorrect or biased inferences of topological relationships. A slower process, *genome rearrangement*, involves fusion of chromosomes from the same or different genomes, irrespectively of the genomic parent. This causes a chromosomal reduction and a ploidy level that differs from the diploid parent. Taking all of the above together, polyploidy is, at least in part, reversible and leads to a smaller genome size over time (Wendel 2015, Leitch 2008).

Recent advances in molecular systematics have revealed the prevalence and the recurrence of paleopolyploid species stemming back to the ancestor of all seed plants (Jiao et al., 2011), with further rounds of genome duplications within various angiosperm lineages (Cui et al., 2006, Soltis et al., 2009, Salmon and Aïnouche, 2015). Due to the common nature of polyploidy in plants, molecular-based data must consider the unpredictable nature of polyploids, such as duplication and loss. Orthologous gene comparisons are required for systematic and evolutionary studies that uses phylogenetic approaches and is therefore of utmost importance (Salmon and Aïnouche, 2015).

Orthologous genes and phylogenies – Orthologous genes descend from a common ancestral DNA sequence (Fig. 2), whereas paralogous genes stem from a gene duplication event (e.g. single gene duplication or genome duplication). Depending on where a speciation or duplication event took place on a phyogeny, it has consequences on the organismal species tree inference. For example, if a speciation occurs after a duplication event then the gene copies 1A and 2A (1B and 2B likewise) are orthologous and trace the origin of speciation between the two species (Fig. 2a). The inheritance of the paralogous gene copies – 1A and 2B or 1B and 2A – would derive from the earlier duplication event, even though the relationships in the gene trees would look the same. For divergence time estimation this will cause a problem because speciation is younger than the duplication event (Fig. 2a). In a different scenario, a duplication event may take place after speciation which makes the duplicated gene copies within each species paralogues (Fig. 2b). The copies within species are paralogues, however, they are orthologues between species because they duplicated after the speciation event (i.e. the common ancestor had one gene shared by both species). It is, therefore, desirable to select many low-copy nuclear genes for species-level phylogenies as it is the only way to reveal orthologues (Álvarez et al., 2008).

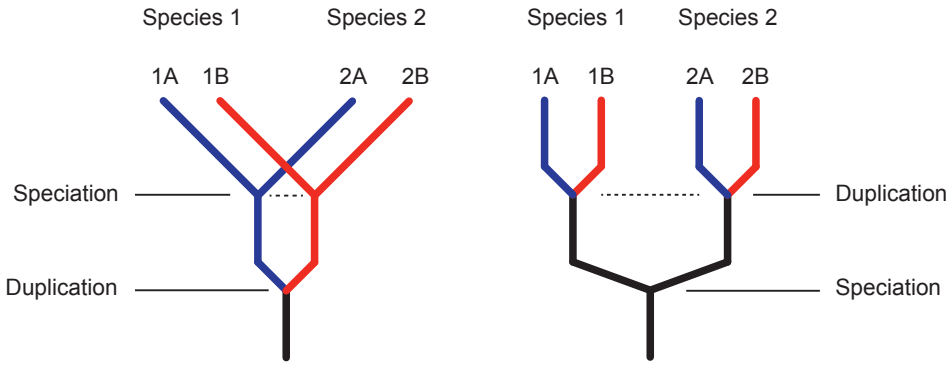


Figure 2. Assessing orthologous and paralogous gene copies affected by duplication. Branches with the same color are orthologues and labeled A and B, respectively.

Orthology and polyploidy – For polyploids, inferring orthology becomes increasingly complex with larger ploidy levels. Orthologous genes can trace back to the speciation of the parental lineages in allopolyploid species (Fig. 1b). However, the duplicated genes within one parental lineage, referred to as paralogues, is tracing the origin of the duplication event and not speciation (Sémon and Wolfe, 2007, Soltis et al., 2010, Salmon and Aïnouche, 2015). Additional complication appears when the progenitors are unknown or have gone extinct as both gene copies would trace the genealogy of the parents (Jones et al., 2013, Marcussen et al., 2014, Bertrand et al., 2015). Thus, the direct evidence of hybrid speciation is obscured with the loss of either parental lineage (Sang, 2002). Therefore, sampling of homoeologues is most likely to be incomplete in paleopolyploids where redundant gene copies may over time be lost to diploidization processes (Marcussen et al., 2014). As a consequence, less data is available to support ancient genome duplication events over recent polyploidizations.

Species tree, networks and reticulate history – Sequencing many single-copy nuclear genes is crucial for understanding polyploid modes. Despite this, reconstructing the reticulate history of polyploids is difficult in a phylogenetic

context. One major difficulty associated with inferring polyploid speciation stems from the fact that polyploids are not accurately represented as a bifurcating or single-labeled tree (Fig. 3). For polyploids and whole genome duplication events (i.e. allopolyploidy) the correct representation of the relationships is a species network or a multi-labeled tree (MUL-tree), where a species can be represented by multiple tips that highlight the inherited homoeologous gene copies from both its parental lineages (Popp and Oxelman, 2001, Marcussen et al., 2014, Gregg et al., 2017). However, tracing the reticulate history of polyploids is not a trivial task and several challenges must be overcome to construct polyploid phylogenies:

I. Retrieving sequences from parental lineages are important. In the absence of known parental lineages the number of genomic combinations increase greatly with the ploidy level. As a consequence, it is often not feasible to reach a single conclusion among the possible hybridization scenarios (Marcussen et al., 2015, Bertrand et al., 2015).
II. Associated with polyploidy is the redundancy of genomic content and the effects of relaxed selection that may lead to deleted copies through pseudogenization (Bertrand 2015). The boundaries for separating orthologs from paralogs and homoeologues from allelic variants may be too small to be successfully recovered (Bertrand et al., 2015).
III. Averaging a species network from a collection of MUL-trees is currently not possible, although advances have been made (e.g. Jones et al., 2013, Marcussen et al., 2014). Methods exist to simplify the minimum number of hybridization events and homoeologous losses using maximum parsimony (e.g. Marcussen et al., 2012). A different approach is to use a diploid backbone tree where each polyploid allele can be analysed individually using bootstrapping to pinpoint the exact location on the backbone tree (Cai et al., 2012). Other methods

construct MUL-consensus tree from gene trees under the assumption that the majority of the genes reflect the genome tree. This consensus tree can then be folded into an uni-labeled network which minimizes the hybridization events (similar to the maximum parsimony method; Huber et al., 2006). However, these methods cannot account for the dubious occurrence of gene duplication (e.g. unrecognized paralogy) or allelic variants (Bertrand et al., 2015), including a method that assigns homoeologues using MCMC under the multispecies coalescent (Jones et al., 2013). Even though the multispecies coalescence method can handle diploid and tetraploid species, it can only assume that hybridizations take place between diploid parental lineages, which is not a strict rule. In addition, the method assumes that the ploidy is known *a priori* (which is not always the case), in order to assign the homoeologues as independent individuals (i.e. independent evolving homoeologue without recombination).

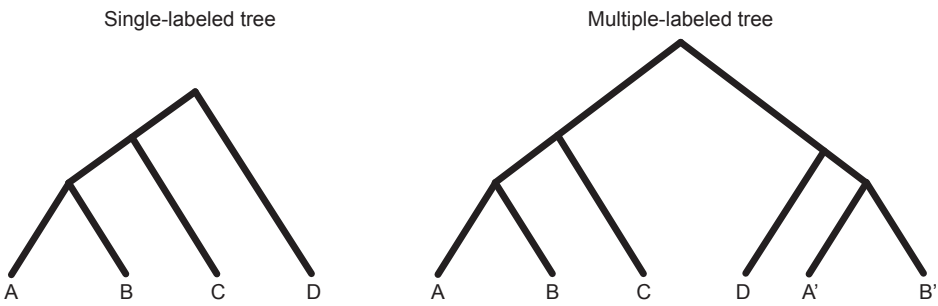


Figure 3. Illustration of single-labeled tree compared to a multiple-labeled tree. On the left-handed side the relationship of four species (A, B, C, D) are illustrated as a single-labeled tree. On the right side, species A and B can be represented by multiple tips (A´ and B´; also known as MUL-tree) that demonstrate their genomic history.

A species tree may also be constructed from MUL-trees by *ad hoc* methods. One way is to treat homoeologues as different individuals (as per Jones et al.,

2013) or remove copies that are due to gene duplications. However, *a priori* information is needed, often as gene trees and species relationships, to correctly identify which copy is orthologous, which is a limiting factor as known ploidy level and chromosome number is often lacking.

Haplotype and allele phasing of HTS data – Separating alleles at a heterozygous gene is important for detecting hybrids (e.g., Eriksson et al., 2018), gene flow (Eriksson et al., 2017, Andermann et al., 2018), population genetics (Martin et al., 2016) and for sampling potential haplotypes needed for understanding the polyploids reticulate history. The numerous haplotypes/homoeologues that can be present in polyploids is a major challenge in bioinformatic applications. For polyploids and gene duplications similar loci may exist in DNA sequence data, which furthermore complicates assessing orthologues, in addition to separating haplotypes, if the parental lineages are close relatives (Bertrand et al., 2015).

Two main methods are used for separating haplotypes; population-based phasing and read-based phasing (i.e. separation of consensus sequence into individual sequences based on variation). The former requires known variants from a reference population where the phasing between alleles is available and can be used to detect new variants. This method is almost exclusively built for model organisms (Kates et al., 2018). For population-based phasing, reads are mapped to a reference (a consensus from a de novo assembly or genome reference) where heterozygous sites are detected. In this approach read depth becomes extremely important for connecting variants as low read depth, often in the intron regions, can result in chimeric haplotypes (Andermann et al., 2018, Eriksson et al., 2018, Kates et al., 2018). Current read-based methods have used different approaches to overcome the challenge of connecting alleles by either using ambiguity codes where the read depth is too shallow to connect two variants (Kates et al., 2018) or using a known pedigree (Browning and

Browning, 2011, Martin et al., 2016). However, the algorithms are built on the assumptions that 1) organisms are diploids and 2) only two haplotypes exist at a locus. These assumptions are violated in the presence of more than two haplotypes, commonly present in neopolyploid and paleopolyploid plants, where current methods may produce either chimeric haplotypes or underestimate the number of haplotypes.

Objectives

The aims of this thesis are to explore the processes that can result in genome duplication, outline the associated challenges in identifying duplications, and identify the implications that genome duplication has on inferring phylogenetic relationships. The work underlying this thesis contained components of field and laboratory work, high-throughput DNA sequencing, bioinformatics, and phylogenetic analyses of polyploid species in plants. This thesis comprises two themes, one reflecting the modes of genome duplication from diploid progenitors (**Paper I & II**) and the other ancient genome duplication that have been obscured by diploidization (**Paper III & IV**). These studies have the following specific objectives:

Paper I – This study investigated the phylogenetic origin of *Medicago prostrata* (Fabaceae), which has multiple ploidy levels. In previous studies of this species, its phylogenetic position varied according to the markers used, suggesting a complicated evolutionary history. We designed a test to distinguish between the biological processes that are known to produce the observed pattern, namely incomplete lineage sorting and hybridization. We tested if *M. prostrata* is a species of hybrid origin or if incomplete lineage sorting can instead explain the observed incongruence among gene trees.

Paper II – In this study we developed a new analytical framework that utilizes high-throughput sequencing data to infer the complex evolutionary history of polyploid taxa. We applied this test to two tetraploid species, *Medicago arborea* and *M. strasseri* (Fabaceae) to investigate if they are of allo- or autopolyploid origin – a question that has remained unanswered despite extensive previous research. Additionally, we tested the efficacy of separating the homoeologous alleles to recover the evolutionary origin of polyploids.

Paper III - In this manuscript we explored if species in Hibisceae display signatures of ancient genome doubling events and whether they are shared by the two whole-genome duplication found in *Hibiscus syriacus.* To do so, we developed a bioinformatic pipeline to separate multiple haplotypes present in polyploids. Three hypothetical genome scenarios were examined using likelihood tests.

Paper IV - The goal of this manuscript was to investigate if the multiple polyploidization events explain the taxonomic uncertainties in *Hibiscus* L. We furthermore selected *Hibiscus* section *Furcaria* to test whether the suggested genome denotations (possible hybridization theories), that had been previously reported through chromosome pairing, could be validated by DNA sequences. Besides numerous nuclear genes, we also sequenced the *RPB2* gene that is known to possess two copies in Hibisceae. We tested whether the common ancestor of Malvoideae (including tribes Gossypieae, Malveae and Hibisceae) had one or two copies of *RPB2* and if the copies in *Hibiscus* had duplicated independently or retained the copies from the ancestor of Malvoideae.

Materials and Methods

Study groups

This thesis includes studies from two lineages of flowering plants, however, the main focus has been on Malvaceae:

Taxonomical classification in Medicago *(Fabaceae)* – The plant genus *Medicago* L. (Fabaceae) consists of 88 described species (Fig. 4; Small 2011, Eriksson et al., 2017) and belongs to the tribe Trifolieae, subtribe Triogonellineae together with *Trigonella* L. and *Melilotus* Mill. (Maureira-Butler et al., 2008). The species are mainly annual and perennial herbs, predominantly found around the Mediterranean basin (Béna et al., 2005) with instances of polyploid species growing in higher latitudes and coastlines (Small, 2011). In section *Dendrotelis*, three species have been found as the only woody sub-shrubs in the genus. The most recent classification established 12 sections and 8 subsections based on the morphological characters in flowers, fruits and seeds (Small and Jomphe, 1989). The phylogenetic relationships are largely unresolved among taxa and there exists a clear incongruence between the morphological characters and molecular delimitation (Maureira-Butler et al., 2008). The observations led to several potential explanations for the incongruence, either due to few markers analyzed or because of several conflicting phylogenetic signals. Some of the markers have low phylogenetic power which is associated with single copy nuclear genes (i.e. there is not enough variations in the deeper nodes) that is used for recovering the true species tree.

Figure 4. *Medicago arborea*. Photo credit Jean Tosti.

Chromosome number variation among taxa – *Medicago* species have undergone multiple polyploidization events in the wild and through cultivation. The tetraploid alfalfa, *M. sativa* L., is one such example of multiple polyploidizations, which is also the world's fourth most economic important crop (Small, 2011). It is primarily used as fodder but also have the potential to become the world's leading source of protein (Small, 2011). The economic importance, together with a low base chromosome number ($x = 8$) and short generation time (three months from germination to frutification), makes *Medicago* species ideal model organisms. For example, the diploid *Medicago truncatula* Gaertn., was the first plant in Fabaceae to be whole-genome sequenced (Young et al., 2011). Besides a complete genome, chromosome counts for most species are available and range from diploids to hexaploids (Small 2011), making them ideal to investigate genome doubling through autopolyploidy (e.g. **Paper I**) and allopolyploidy (e.g. **Paper II**).

Genome duplication events – Analysis of plant genomes recognized a shared whole-genome triplication event preceding the rosid-asterid split (Young et al., 2011, Wendel, 2015). A second genome duplication was strongly suggested to have taken place in the legumes approximately 58 Mya (Pfeil et al., 2005).

Taxonomical classification in Malvaceae-- Malvaceae is a cosmopolitan family and mainly concentrated in the tropical regions (La Duke and Doebley, 1995). Current taxonomic classification divides Malvaceae into three tribes: Gossypieae, Hibisceae and Malveae (Pfeil and Crisp, 2005, Koopman and Baum, 2008). In particular, Hibisceae is taxonomically challenging due to morphological characters shared with the other tribes while lacking synapomorphies within the tribe that would establish the evolutionary relationships among the species (Pfeil et al., 2002). *Hibiscus* L. is one such example where the ambiguous circumscription stems from a conflict between distinctive morphological characters (Pfeil and Crisp, 2005) and the clear, but different, nuclear and chloroplast gene results, causing a paraphyletic genus *Hibiscus* (Pfeil et al. 2002, Pfeil and Crisp, 2005, Koopman and Baum 2008). Four informally named subclades are now recognized – i.e., /Phylloglandula (e.g., *Hibiscus* sect. *Furcaria* (Fig. 5; /Furcaria), *Urena*, *Decaschistia*), /Trionum (*H.* sect. *Trionum*, *Abelmoschus*, *Pavonia* and *Malvaviscus*), /Euhibiscus (e.g., *H. rosa-sinensis, H. syriacus*) and /Calyphylli (e.g.; *H. calyphyllum*) – instead of establishing a large number of genera or synonymizing everything into *Hibiscus* (Pfeil and Crisp, 2005). A fifth subclade, /Megistohibiscus, sister to the rest was recognized by Koopman & Baum (2008). The previously recognized tribe Malvavisceae (now included in /Trionum) presents the greatest challenge to the traditional classification of *Hibiscus*. Namely that the largest subclade – *Pavonia* Cav. with over 220

species – is the least resolved and understood group of species in *Hibiscus* (Fryxell, 1999, Pfeil and Crisp, 2005).

Figure 5. *Hibiscus radiatus*. Photo credit Prenn (https://commons.wikimedia.org/wiki/File:Hibiscus_radiatus_01.JPG).

Chromosome number variation -- The base chromosome numbers within Malvoideae varies greatly between subclades (e.g. *Gossypium* n = 13; *Malva* = 7, /Trionum n = 7 or 14, /Furcaria n = 18 and /Euhibiscus n = 20-22; Bates, 1967, Bates and Blanchard Jr, 1970, Fryxell, 1999) and species show high diversity of chromosome numbers within genera, often in exact multiples of base numbers (e.g. /Furcaria with 2n = 36, 72, 144). This pattern suggests an explanation of several rounds of increased ploidy level. In particular, thirteen different diploid genomes have been identified in clade /Furcaria through hybridization experiments (Menzel, 1966, Menzel and Wilson, 1969; Fig. 6), providing evidence about the origins of polyploid taxa, resulting in polyploid offspring, tetraploid to decaploids, resembling extant species (Menzel, 1966, Menzel and Wilson, 1969). The Australian alliance of /Furcaria (e.g. *H.*

heterophyllus) are the only natural hexaploids known in the section and cytotaxonomical studies have suggested that all species from the Australian alliance stem from allohexaploidy (Menzel and Martin, 1974) with further speciation.

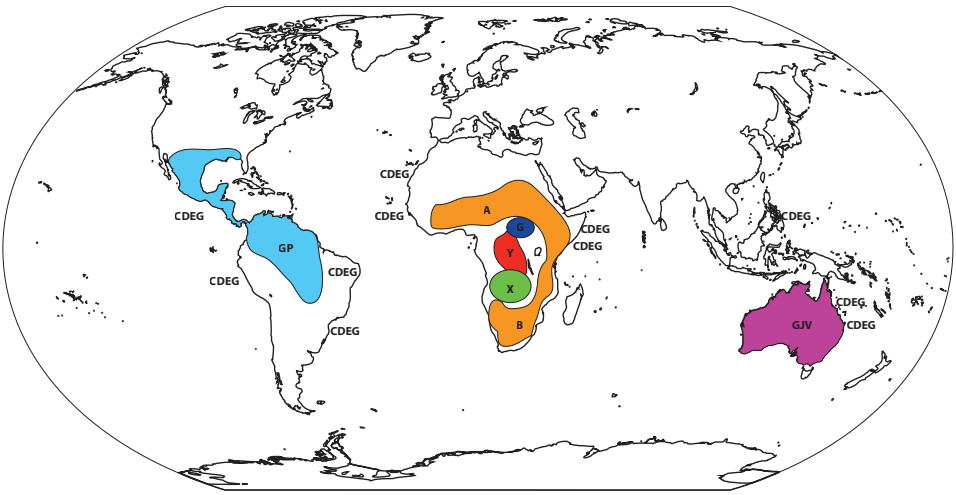


Figure 6. Genome distributions in *Hibiscus* section *Furcaria*. The diploid genomes are mainly distributed in Africa. The tetraploids occur in Africa, Asia, America and Australia. The hexaploid alliance (GJV; i.e. *H. heterophyllus*) occurs only in Australia. The octoploid (CDEG; *H. diversifolius*) has a cosmopolitan distribution.

Genome duplication events – Beside recent formations of polyploidization (i.e. extant species formed through auto- and allopolyploidy), ancient genome duplications have been recognized in Malvoideae (e.g. Paper III & IV; Wang et al., 2012, Kim et al., 2017). Genome data from the diploid *Gossypium raimondii* Ulbr. established two polyploidization events, where one is shared by the eudicots (Wang et al., 2012). Similarly has *Hibiscus syriacus* L. been shown to have undergone at least one independent genome doubling within the lineage (Kim et al., 2017), although, evidence of a second event was confirmed to be shared by all species in *Hibiscus* (**Paper III**). Considering the diverse

base chromosome numbers and high ploidy levels in extant species of *Hibiscus*, likely rounds of polyploidizations may exist among the subclades.

DNA extraction and sequencing

Plant material and DNA extraction – Plant materials were obtained as seeds grown in plant growth chambers at the University of Gothenburg, and from cuttings sourced from the Gothenburg Botanical Garden. Seeds for **Paper I & II** came from the United States Department of Agriculture (USDA) and from Sienna Botanic Garden, Italy. All seeds were soaked with a small amount of detergent for 24 h before sowing (further detail in Sousa, 2015). For **Paper III & IV** seeds and cuttings were obtained from the USDA, Botanic Garden Meise (BR), Botanische Gärten der Universität Bonn (BONN), and the Royal Botanic Gardens Kew (K). *Hibiscus* seeds are protected by hard mericarp walls that was cut with a razorblade before sowing in pots. After 2-3 days of germination, one seedling was kept per pot and grown in chambers until they flowered and reached frutification. DNA material was collected as leaves that were dried using silica gel. Vouchers were made and stored at Gothenburg herbarium (GB) after DNA material were collected. In addition to silica dried material, leaf tissue from herbarium specimens were obtained from herbaria GB, K, and BR. DNA extraction was performed using DNeasy Plant mini Kit (Qiagen, Valencia, CA, USA) under standard protocols. Samples that had excessive polysaccharides (high sugar content) and phenols (colors) required an additional cleaning step with 99% ethanol.

High-throughput sequencing - DNA sequences were generated from high-throughput sequencing using the Illumina MiSeq platform (San Diego, California, USA) in **Papers I-IV**. The advantage of using high-throughput sequencing compared to Sanger sequencing protocols lies in the vast amount of data and the number of genes targeted for a considerable reduction in price

per nucleotide. High-throughput sequencing technique often includes enough sequence data to identify species of cryptic origin (**Paper I**), species that possess genomes from multiple origins (**Paper II, IV**), and ancient genome doubling in the presence of diploidization (**Paper III**). In contrast, Sanger sequencing requires primer design whenever a gene has been found to possess two copies. A considerable amount of time is saved removing the additional steps included for designing internal primers, even though high-throughput sequencing requires a more complicated lab protocol. Despite the advantages of high-throughput sequence technique, it introduces new challenges such as how to handle the large amount of data and lack of standardized bioinformatic protocols.

Methodological challenges – In the pipeline developed for **Papers I-II**, we found problems of merging contigs – a set of overlapping reads that together represent a consensus region – produced by the phasing tool, whenever more than two copies were present or because of a low read depth in the introns. The issue was resolved by merging the contigs and testing for recombination. In the absence of paralogous copies, formed through gene or genome duplication, this approach worked well, however this was not applicable in Malvaceae (**Papers III-IV**). Dealing with tetraploids to octoploids, multiple contigs were produced by the phasing tool and merging these contigs by hand was not feasible. Even testing the merged contigs with a recombination program (e.g. RDP4; Martin et al., 2015) was unsuccessful due to the numerous combinations of possible haplotypes. Furthermore, available pipelines failed to identify all possible haplotypes after checking the output with the data that had not been phased. Since current phasing tools are developed to separate only two haplotypes at a single locus, they only separate two copies. These two copies could therefore be the result of several haplotypes that were phased at random, mixing the haplotypes rather than separating them. Papers **III-IV**

describe how the issues of separating haplotypes were overcome, without limiting the pipeline to assume diploid species, single-gene copy or only separating two haplotypes.

Results and Discussion

The chapters in this thesis explore some of the complexities surrounding polyploid species. They bring up issues concerning potential cryptic species in polyploid taxa (**Paper I**) and the difficulty to interpret the mode of origin without extensive data (**Paper I-II**). In both these papers, the high number of loci, taxon sampling and data made it possible to unravel some of the long-standing questions identified in previous studies. Furthermore, the results also demonstrate the importance of careful curation of high-throughput sequence data in the presence of ancient genome duplication and multiple haplotypes (**Paper II-IV**).

The main results and conclusions of each paper are presented below:

Paper I – In this paper we described a new species, *Medicago tetraprostrata* J.S. Erikss. & B.E. Pfeil, from the observations of two chromosome counts and the support of separate evolutionary lineages between the diploid and tetraploid individuals. It is not uncommon to find diploid and tetraploid chromosome counts in *Medicago* species, in fact there are six described species exhibiting this phenomenon (Small, 2011). By examining eight nuclear genes from 22 individuals of *M. prostrata*, we assessed the ploidy level by separating the alleles for each individual where some were diploids and others tetraploids. We confirmed our assessment by comparing the sequenced-based inference of ploidy level with chromosome counts of one diploid and one tetraploid individual from the same dataset. With the reconstructed gene trees, we inferred the evolutionary relationships of the diploid and tetraploid individuals. The diploids were not supported as being of homoploid hybrid origin (hybridization without genome duplication). Instead, we suggested that *M. tetraprostrata* is of autopolyploid origin with some introgression of alleles from other tetraploid species (mainly from the *M. sativa* complex; Fig. 7).

From these results we concluded that the tetraploid individuals constitute a cryptic species and should be recognized as a separate taxonomic entity from its diploid progenitor.

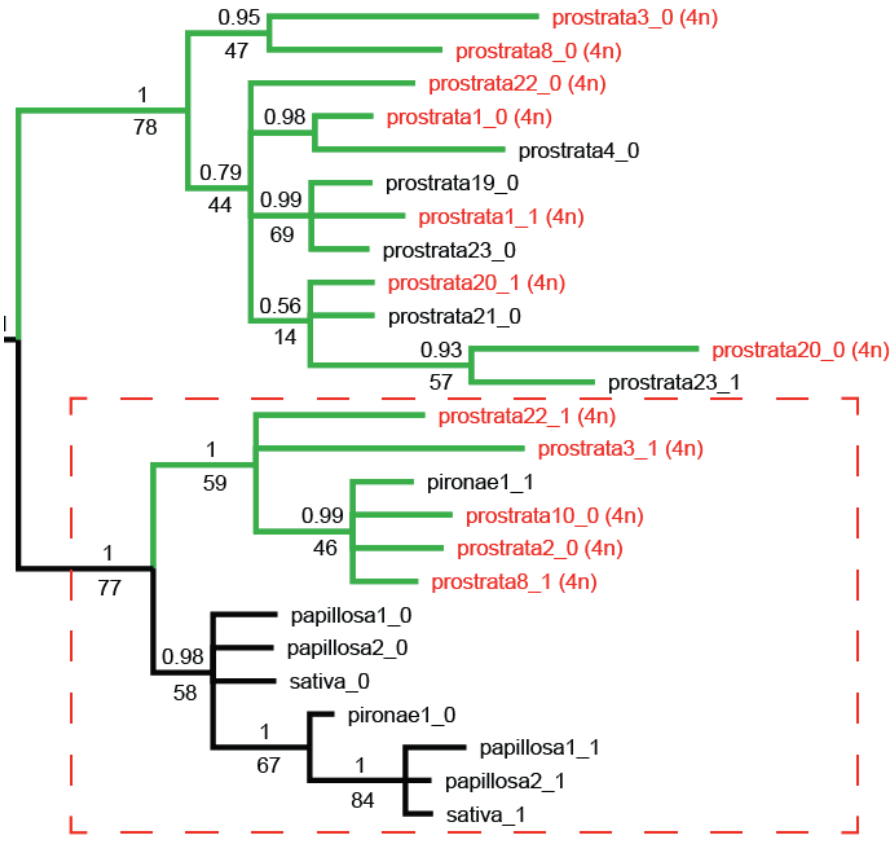


Figure 7. Autopolyploid gene tree with introgression. Modified figure for illustrating the two ploidy groups position in the gene tree (from figure 3 in Paper I). Red text labels specify the tetraploid individuals in the study. The box around the Sativa clade demonstrate an instance of introgression with the tetraploids. Posterior probability above branches and bootstrap below.

Paper II – In this paper we found that separating potentially heterozygous loci into separate alleles was crucial to understanding the polyploid mode of origin (i.e. auto- or allopolyploidy). In two tetraploid species, *Medicago arborea* and *M. strasseri*¸ the origin was unresolved in previous cytological studies (Rosato et al., 2008). Genetic data supported the previously reported chromosome counts from cytological studies that both species are tetraploids. Allele phasing accurately inferred the polyploid mode of origin. Whenever using the majority rule consensus sequences (by selecting the nucleotide that occurred most frequently in the read data) the overall phylogenetic resolution was lower, obscuring any possible signal of polyploid mode of origin. In eight genes out of ten, we found that homoeologues separated into two distinct clades sister to either potential diploid genome donor, a pattern suggesting allopolyploidy (Fig. 8, left-handed tree). In one gene tree the two clades were sisters but distinct from each other, a pattern typical of autopolyploidy (Fig. 8, right-handed tree). Taken together, we determined that species of *Dendrotelis* originated through hybridization with subsequent genome duplication (i.e. allopolyploidy). The minority pattern, however, seen in one gene where the alleles formed one clade (autopolyploidy) could not be dismissed. A mix between alleles forming one or two separate clades sister to other taxa may be evidence of *segmental allopolyploidy*. This form of genome duplication is possible if two closely related species hybridized and parts of their genomes are similar enough that chromosomes could recombine between parental genomes (i.e. the definition of autopolyploidy). Lastly, our results show that woodiness in *Dendrotelis* is a derived morphological trait from herbaceous ancestors (Steele et al., 2010). We suggest that woodiness may be a transgressive phenotype, caused by polyploid hybridization.

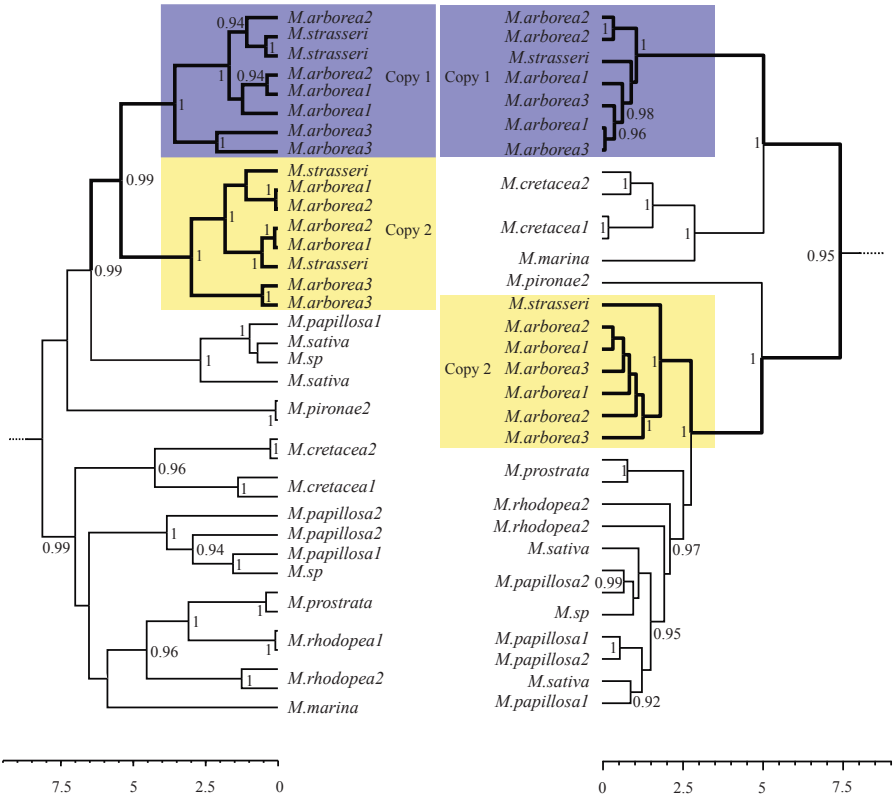
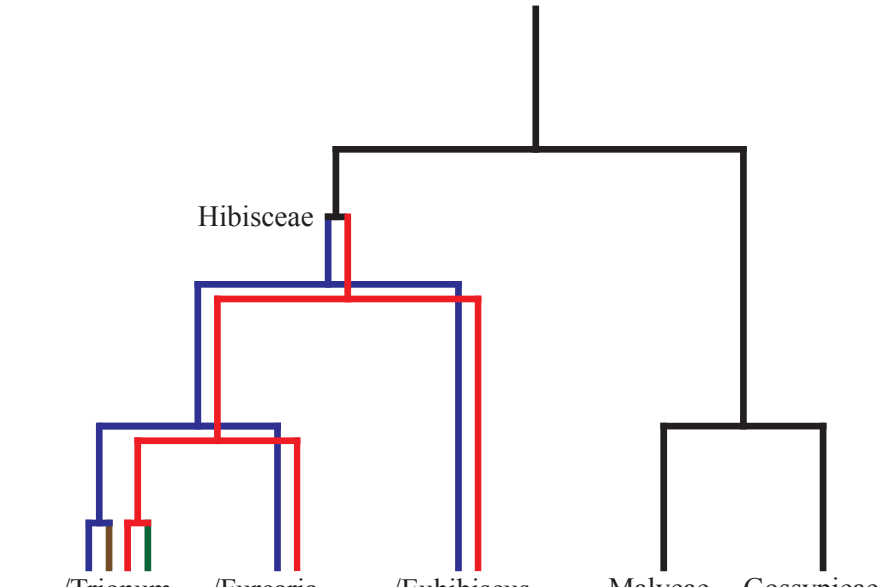


Figure 8. Two gene trees illustrating both modes of polyploidy found in two species. The two-colored boxes have sequence copies represented by all individuals separated as two clades. In the left-handed tree, the two clades form a monophyly, representing an autopolyploid origin, whereas the right-handed tree the two clades are sister to respective parental lineages, characteristic of allopolyploid origin.

Paper III – It is widely accepted that polyploids have complex evolutionary histories, however diploids may have equally complicated patterns. In this manuscript we present evidence that several whole-genome duplication events took place in *Hibiscus* (Malvaceae). Numerous single-copy nuclear genes were found to be paralogues in the diploid species *H. cannabinus* and *H. mechowii* from *Hibiscus* sect. *Furcaria*, supporting a scenario that duplicated genes were retained after diploidization. These paralogues could not be explained by single

gene duplications using likelihood model-based scenario testing. Instead the best-supported scenario defined two whole-genome duplication events (one independent genome duplication leading to *H. syriacus* and one genome duplication shared by all species of *Hibiscus*) that shaped the Hibisceae genome. The two previously reported genome duplications in *H. syriacus* (Kim et al. 2017) were corroborated in this study, although with one exception; one of the duplication events is older than previously understood (Kim et al., 2017) and occurred somewhere along the branch leading to the clade *Hibiscus*. Additionally, during the process of retrieving haplotypes using the newly developed pipeline it was revealed that species from clade /Trionum (i.e. represented by *Pavonia triloba* and *H. trionum*) always possessed two haplotypes, whereas species of *Pavonia* possessed twice as many copies regardless of they being considered paralogous or single-copy genes. The additional haplotypes in /Trionum led to a third genome duplication that was independent, in addition to the two genome duplications found in *Hibiscus* (Fig. 9). These results were only retrieved from the Illumina data by creating a new allele phasing pipeline that does not assume diploidy with only two haplotypes at a locus. This pipeline can be adapted to polyploids or non-polyploids alike, paralogous or single-copy genes, without being exclusive to model organisms. Considering the diverse chromosome numbers in plants, more evidence of ancient genome duplications and processes of diploidizations are yet to be explored.



Base chromosome number				
n = 7-14	n = 18	n = 20-22?	n = 6-8	n = 13

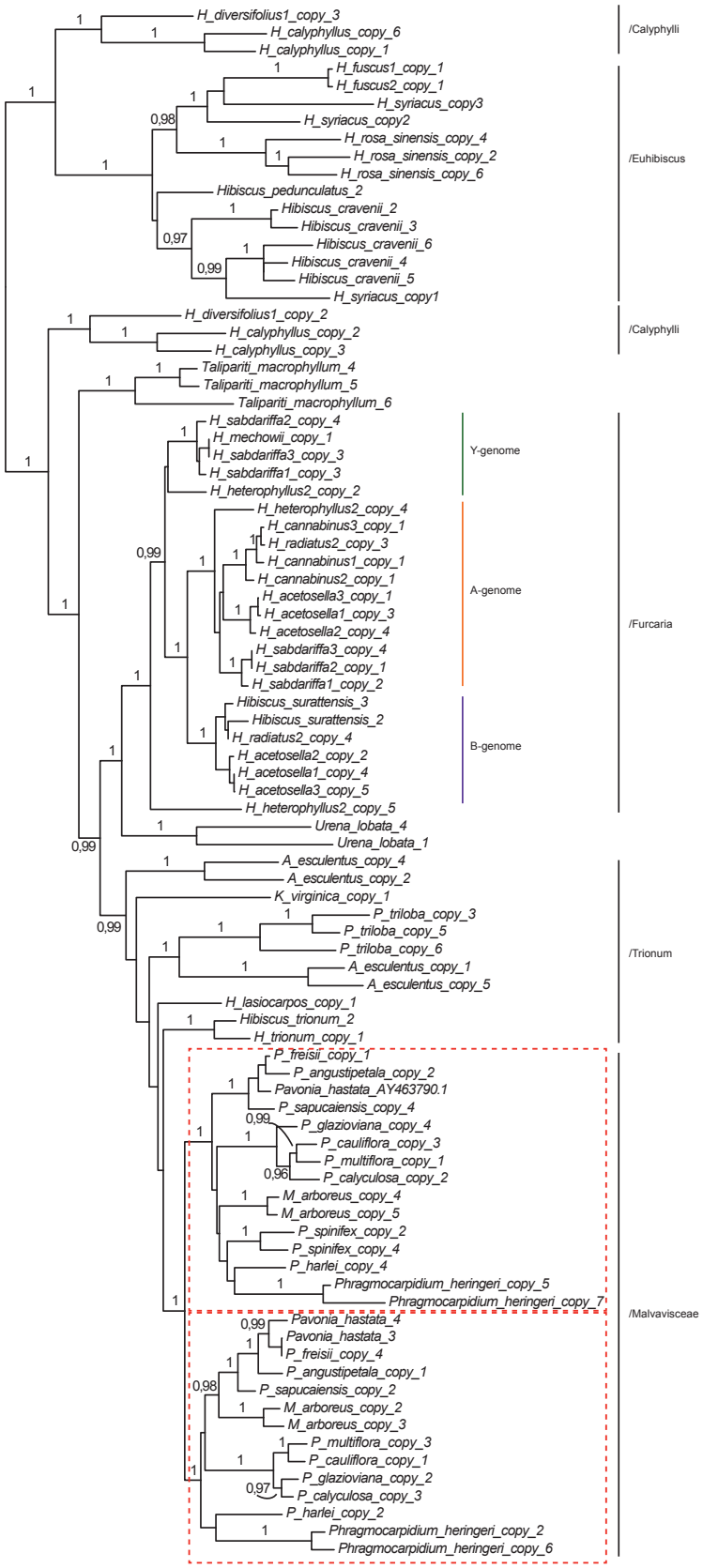
Figure 9. Updated genome evolution of subfamily Malvoideae (Malvaceae). The colored branches indicate the genome duplications that have been established in paper III. The base chromosome number for each subclade (/Trionum, /Furcaria and /Euhibiscus) and tribes (Gossypieae and Malveae) are found under respective tips.

Paper IV – In this study we found that the numerous sequence copies of *RPB2* in our data reflect two independent genome duplication events in Malvaceae. One of the duplications predated the divergence of Hibisceae and the other predated the divergence of /Trionum. Furthermore, one long-standing question – whether the ancestor of Malvaceae possessed one or two *RPB2* copies – was answered in this study. The occurrence of a single *RPB2-d* copy in the ancestor of Malvaceae is supported by two lines of evidence: a single occurrence of *RPB2-d* copy in *Theobroma* and *Herrania*, sister to Malvoideae and the tribes Hibisceae, Gossypieae and Malveae; and the *RPB2* copies found in Hibisceae formed a monophyletic group within the tribe. We also found no evidence for

the ancestor of Malvoideae having two copies, with a copy going extinct in the lineage to modern-day Malveae and Gossypieae.

The phylogenetic gene trees corroborate the denoted genome-labels established by Menzel (1966-1969) in /Furcaria. The tetraploid hybrids appear next to either parental A-, B-, or Y-genome lineage. Furthermore, the B- and Y-genomes share a common ancestor. Species in the /Trionum clade were also discovered to possess additional RPB2 copies forming two monophyletic groups sister to respective /Furcaria paralogue of RPB2 (Fig. 10). The reason behind the challenging taxonomical classification of *Hibiscus* is validated by the frequent polyploidizations among closely related species and genera, together with the varying haploid chromosome numbers for each subclade. *Hibiscus* is still paraphyletic in its standing condition, however, utilizing high-throughput sequence data with chromosome counts, a clearer picture of the complex reticulate history can be made. The large subclade /Malvavisceae, is also affected by polyploidization events involving different genera, resulting in an even greater taxonomical mess than *Hibiscus*.

Figure 10, next page. Modified gene phylogeny of one of the two RPB2 copies in Hibisceae. Subclades are shown on the right side of the tree. Clade /Furcaria is further defined into three genomic clades A-, B- and Y-genomes, each with one diploid species as placeholder (e.g. *H. cannabinus* [A], *H. surattensis* [B] and *H. mechowii* [Y]). Red boxes demonstrate a duplication event involving several species that appear in both clades.


H_diversifolius1_copy_3
H_calyphyllus_copy_6
H_calyphyllus_copy_1
/Calyphylli
H_fuscus1_copy_1
H_fuscus2_copy_1
H_syriacus_copy3
H_syriacus_copy2
H_rosa_sinensis_copy_4
H_rosa_sinensis_copy_2
H_rosa_sinensis_copy_6
/Euhibiscus
Hibiscus_pedunculatus_2
Hibiscus_cravenii_2
Hibiscus_cravenii_3
Hibiscus_cravenii_6
Hibiscus_cravenii_4
Hibiscus_cravenii_5
H_syriacus_copy1
H_diversifolius1_copy_2
H_calyphyllus_copy_2
H_calyphyllus_copy_3
/Calyphylli
Talipariti_macrophyllum_4
Talipariti_macrophyllum_5
Talipariti_macrophyllum_6
H_sabdariffa2_copy_4
H_mechowii_copy_1
H_sabdariffa3_copy_3
H_sabdariffa1_copy_3
H_heterophyllus2_copy_2
Y-genome
H_heterophyllus2_copy_4
H_cannabinus3_copy_1
H_radiatus2_copy_3
H_cannabinus1_copy_1
H_cannabinus2_copy_1
H_acetosella3_copy_1
H_acetosella1_copy_3
H_acetosella2_copy_4
H_sabdariffa3_copy_4
H_sabdariffa2_copy_1
H_sabdariffa1_copy_2
A-genome
Hibiscus_surattensis_3
Hibiscus_surattensis_2
H_radiatus2_copy_4
H_acetosella2_copy_2
H_acetosella1_copy_4
H_acetosella3_copy_5
B-genome
H_heterophyllus2_copy_5
/Furcaria
Urena_lobata_4
Urena_lobata_1
A_esculentus_copy_4
A_esculentus_copy_2
K_virginica_copy_1
P_triloba_copy_3
P_triloba_copy_5
P_triloba_copy_6
A_esculentus_copy_1
A_esculentus_copy_5
/Trionum
H_lasiocarpos_copy_1
Hibiscus_trionum_2
H_trionum_copy_1
P_freisii_copy_1
P_angustipetala_copy_2
Pavonia_hastata_AY463790.1
P_sapucaiensis_copy_4
P_glazioviana_copy_4
P_cauliflora_copy_3
P_multiflora_copy_1
P_calyculosa_copy_2
M_arboreus_copy_4
M_arboreus_copy_5
P_spinifex_copy_2
P_spinifex_copy_4
P_harlei_copy_4
Phragmocarpidium_heringeri_copy_5
Phragmocarpidium_heringeri_copy_7
Pavonia_hastata_4
Pavonia_hastata_3
P_freisii_copy_4
P_angustipetala_copy_1
P_sapucaiensis_copy_2
M_arboreus_copy_2
M_arboreus_copy_3
P_multiflora_copy_3
P_cauliflora_copy_1
P_glazioviana_copy_2
P_calyculosa_copy_3
P_harlei_copy_2
Phragmocarpidium_heringeri_copy_2
Phragmocarpidium_heringeri_copy_6
/Malvavisceae

Conclusions

The high frequency of plant polyploidy implies that all aspects of research must take the polyploid nature of plant genomes into account (Salmon and Aïnouche, 2015). In this thesis several challenges associated with polyploid evolution in plants were identified, new laboratorial and methodological advances were presented to tackle these challenges. I specifically focused on the difficulties to infer the mode of origin linked with the increase of genome size, e.g. polyploidization; the ongoing processes of polyploidization and diploidization that lead to a fragmented genome; the methodological limitations to capture the traces of polyploid origin in plants. These studies focused on plant families that are known to contain polyploid taxa, but whose processes leading to the variation in genome size and chromosome numbers were insufficiently known.

Plant taxa have frequently been reported with several ploidy levels and base chromosome numbers. This thesis contributed evidence that in one such case, rather than a miscount of chromosomes, a cryptic species originated from a duplication of the genome (**Paper I**). When morphological characters are inconspicuous and chromosome counts may be misleading, if it is due to miscounts, genetic data such as target capture enrichment selecting for many genes can uncover the "cryptic" species nested in a taxonomically legitimate name. The findings in **Paper I** suggest that polyploid formation through autopolyploidy is not an evolutionary dead-end (Wagner Jr, 1970), but may in fact be a bridge for gene flow between species.

An important result with implications for the broad plant phylogenetic community is that even diploid plants show evidence of whole-genome duplications. Selecting for single-copy genes may as a default obscure any signals of duplication events, if the genes have undergone copy loss through

the diploidization process that return a polyploid genome to a diploid-like state. Hence, selecting for genes to elucidate the evolution of plant taxa becomes extremely important: as **Paper III** demonstrates, even single-copy genes are likely to be paralogues. The large variation in base chromosome numbers in Malvaceae indicate how complicated genome evolution may be in different tribes and genera (Fig. 9). This study provided new insights into genome duplication, polyploidy and diploidization processes that have shaped species evolution in Malvaceae.

With the increased use of high-throughput sequencing data, which is an enormous advantage for polyploid studies, this thesis has also demonstrated the need for methodological advances working with any ploidy species, whether it is a diploid or a polyploid. The generation and identification of all possible haplotypes are very important for recognizing the mode of polyploidization and the loss of genes through diploidization. Although, the strict threshold for accepting potential haplotypes in the pipeline (**Paper III-IV**) may have underestimated "true" sequence copies (i.e. rare haplotypes) that were removed because of low support. Even with the advances of sequencing techniques, I recommend that chromosome counts should continue to be established in plant taxa as it provides a broader context to genome evolution and the accompanying processes.

Future prospects

Polyploidization has been demonstrated in this thesis to be a major driver of speciation in two distantly related families, however, many questions still remain to be answered. Neo- and paleopolyploidization events were confirmed with the generated high-throughput sequence data, validating that enough data can be produced through target gene capture and without the need for whole-genome data. In Malvaceae, and especially regarding *Hibiscus,* further sampling of species is needed to fully understand the potential polyploidization events that may cause the currently disputed classification. Although, not to be forgotten, additional cytotaxonomical studies should be continued to establish a foundation of background information that is useful for connecting the molecular work with the biological processes. This is especially important for higher polyploids where gene copies have a higher chance to be lost over time (e.g. gene conversion), and the difficulty of capturing rare alleles.

On the other hand, sequence data produced from target gene capture cannot reveal chromosome evolution. *Hibiscus* is an example where clades, /Furcaria, /Euhibiscus and /Trionum have three base chromosome numbers that differ between the clades. A selection of genes is often unable to target the positions where potential fusion/fission of chromosomes has occurred. Therefore, acquiring genome sequence data and identifying genomic blocks may reveal the evolution of chromosomes and how they have resolved into three different base chromosome numbers in *Hibiscus*.

Lastly, what to do with *Hibiscus*? It is evident that polyploidy is a common process in this group and that hybridizations have driven the speciation. Ongoing research is prying into the real mess of *Hibiscus*, focusing on different groups that are known to be taxonomically complex (e.g. *H. trionum* possessing several ploidy levels) and species which only occurs as hexaploids,

like the Australian /Furcaria alliance). Sampling more specimens, validating their ploidy level, and using the same genetic markers will open up for adding new information on existing material. This has been done between this project and the Australian group of *Hibiscus* species (section *Trichospermum*), which may be joined with the same genetic markers. *Hibiscus* continues to be unsolved even after this study, however, a good foundation to continue this work has been established.

Paper contributions

Paper I – *A cryptic species produced by autopolyploidy and subsequent introgression involving* Medicago prostrata *(Fabaceae).* Jonna Sofia Eriksson (JSE) with Bernard E. Pfeil (BEP) conceived this study together with the support from co-authors. JSE and Filipe de Sousa (FdS) carried out all the lab work jointly. JSE, BEP performed all the analyses together with co-authors. FdS and Yann J. K. Bertrand (YJKB) wrote the code. JSE and BEP wrote the manuscript with contributions from all co-authors.

Paper II – *Allele phasing is critical to revealing a shared allopolyploid origin of* Medicago arborea *and* M. strasseri *(Fabaceae).* JSE with BEP conceived this study together with the support from co-authors. JSE carried out all the lab work and performed all the analyses with the support from co-authors. JSE and BEP wrote the manuscript with contributions from all co-authors.

Paper III – *Two ancient genome duplication events shape diversity in Hibiscus L. (Malvaceae).* JSE conceived this study with the support from co-authors. JSE carried out all the field and laboratory work, and performed all the analyses with support from co-authors. JSE and Dominic John Bennett (DJB) wrote the code. JSE led the writing with contributions from all co-authors.

Paper IV – *Base chromosome number variations and major polyploidization events impact taxonomic classification in Hibisceae.* JSE conceived this study with the support from co-authors. JSE carried out all the field and laboratory work, performed all the analyses with support from co-authors. DJB wrote the code. JSE led the writing with contributions from all co-authors.

References

ÁLVAREZ, I., COSTA, A. & FELINER, G. N. 2008. Selecting Single-Copy Nuclear Genes for Plant Phylogenetics: A Preliminary Analysis for the Senecioneae (Asteraceae). *Journal of Molecular Evolution,* 66**,** 276-291.

ANDERMANN, T.*, et al.* 2018. Allele phasing greatly improves the phylogenetic utility of ultraconserved elements. *Systematic biology***,** syy039-syy039.

BATES, D. M. 1967. Chromosome number in the Malvales. I. *Gentes Herb.,* 10.

BATES, D. M. & BLANCHARD JR, O. J. 1970. Chromosome numbers in the Malvales. II. New or otherwise noteworthy counts relevant to classification in the Malvaceae, tribe Malveae. *American Journal of Botany***,** 927-934.

BÉNA, G., LYET, A., HUGUET, T. & OLIVIERI, I. 2005. Medicago – Sinorhizobium symbiotic specificity evolution and the geographic expansion of Medicago. *Journal of Evolutionary Biology,* 18**,** 1547-1558.

BERTRAND, Y. J.*, et al.* 2015. Assignment of Homoeologs to Parental Genomes in Allopolyploids for Species Tree Inference, with an Example from Fumaria (Papaveraceae). *Systematic biology,* 64**,** 448-471.

BLANC, G. & WOLFE, K. H. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The plant cell,* 16**,** 1667-1678.

BROWNING, S. R. & BROWNING, B. L. 2011. Haplotype phasing: Existing methods and new developments. *Nature Reviews. Genetics,* 12**,** 703-714.

CAI, D.*, et al.* 2012. Single copy nuclear gene analysis of polyploidy in wild potatoes (Solanum section Petota). *BMC evolutionary biology,* 12**,** 70.

CUI, L.*, et al.* 2006. Widespread genome duplications throughout the history of flowering plants. *Genome research,* 16**,** 738-749.

DOYLE, J. J.*, et al.* 2008. Evolutionary genetics of genome merger and doubling in plants. *Annual review of genetics,* 42**,** 443-461.

ERIKSSON, J.*, et al.* 2017. A cryptic species produced by autopolyploidy and subsequent introgression involving Medicago prostrata (Fabaceae). *Molecular phylogenetics and evolution,* 107**,** 367-381.

FRYXELL, P. A. 1999. Pavonia Cavanilles (Malvaceae). *Flora neotropica monograph,* 76.

GRANT, V. 1981. *Plant speciation*, New York: Columbia University Press.

GREGG, W. C. T., ATHER, S. H. & HAHN, M. W. 2017. Gene-Tree Reconciliation with MUL-Trees to Resolve Polyploidy Events. *Systematic Biology,* 66**,** 1007-1018.

HAVANANDA, T., BRUMMER, E. C. & DOYLE, J. J. 2011. Complex patterns of autopolyploid evolution in alfalfa and allies (Medicago sativa; Leguminosae). *American journal of botany,* 98**,** 1633-1646.

HUBER, K. T., OXELMAN, B., LOTT, M. & MOULTON, V. 2006. Reconstructing the evolutionary history of polyploids from multilabeled trees. *Molecular Biology Evolution,* 23**,** 1784-1791.

JIAO, Y.*, et al.* 2012. A genome triplication associated with early diversification of the core eudicots. *Genome biology,* 13**,** R3.

JIAO, Y., *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature,* 473**,** 97.

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.

JUDD, W. S. & MANCHESTER, S. R. 1997. Circumscription of Malvaceae (Malvales) as determined by a preliminary cladistic analysis of morphological, anatomical, palynological, and chemical characters. *Brittonia,* 49**,** 384-405.

KATES, H. R., *et al.* 2018. Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of Artocarpus. *American journal of botany,* 105**,** 404-416.

KIM, Y., *et al.* 2017. Genome analysis of Hibiscus syriacus provides insights of polyploidization and indeterminate flowering in woody plants. *DNA Research,* 24**,** 71-80.

KOOPMAN, M. M. & BAUM, D. A. 2008. Phylogeny and biogeography of tribe Hibisceae (Malvaceae) on Madagascar. *Systematic Botany,* 33**,** 364-374.

LA DUKE, J. C. & DOEBLEY, J. 1995. A chloroplast DNA based phylogeny of the Malvaceae. *Systematic Botany***,** 259-271.

MARCUSSEN, T., *et al.* 2014. From gene trees to a dated allopolyploid network: insights from the angiosperm genus Viola (Violaceae). *Systematic biology,* 64**,** 84-101.

MARCUSSEN, T., *et al.* 2012. Inferring species networks from gene trees in high-polyploid North American and Hawaiian violets (Viola, Violaceae). *Systematic biology,* 61**,** 107-126.

MARTIN, D. P., *et al.* 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus evolution,* 1.

MARTIN, M., *et al.* 2016. WhatsHap: fast and accurate read-based phasing. *bioRxiv***,** 085050.

MARTINO, A. L. & SINSCH, U. 2002. Speciation by polyploidy in Odontophrynus americanus. *Journal of Zoology,* 257**,** 67-81.

MAUREIRA-BUTLER, I. J., *et al.* 2008. The reticulate history of Medicago (Fabaceae). *Systematic Biology,* 57**,** 466-482.

MENZEL, M. Y. 1966. The pachytene chromosome complement of Hibiscus cannabinus. *Cytologia,* 31**,** 36-42.

MENZEL, M. Y. & MARTIN, D. W. 1974. Cytotaxonomy of some Australian species of Hibiscus sect. Furcaria. *Australian Journal of Botany,* 22**,** 141-156.

MENZEL, M. Y. & WILSON, F. D. 1969. Genetic relationships in hibiscus sect. Furcaria. *Brittonia,* 21**,** 91.

OHNO, S. 1970. Evolution by gene duplication. London. George Allen and Unwin.

ORR, H. A. 1990. " Why polyploidy is rarer in animals than in plants" revisited. *J The American Naturalist,* 136**,** 759-770.

OTTO, S. P. & WHITTON, J. 2000. Polyploid incidence and evolution. *Annual review of genetics,* 34**,** 401-437.

PELLICER, J., HIDALGO, O., DODSWORTH, S. & LEITCH, I. J. 2018. Genome size diversity and its impact on the evolution of land plants. *Genes,* 9**,** 88.

PFEIL, B., BRUBAKER, C., CRAVEN, L. & CRISP, M. 2002. Phylogeny of Hibiscus and the tribe Hibisceae (Malvaceae) using chloroplast DNA sequences of ndhF and the rpl16 intron. *Systematic Botany***,** 333-350.

PFEIL, B. & CRISP, M. 2005. What to do with Hibiscus? A proposed nomenclatural resolution for a large and well known genus of Malvaceae and comments on paraphyly. *Australian Systematic Botany,* 18**,** 49-60.

PFEIL, B. E., SCHLUETER, J. A., SHOEMAKER, R. C. & DOYLE, J. J. 2005. Placing Paleopolyploidy in Relation to Taxon Divergence: A Phylogenetic Analysis in Legumes Using 39 Gene Families. *Systematic Biology,* 54**,** 441-454.

POPP, M. & OXELMAN, B. 2001. Inferring the History of the Polyploid Silene aegaea (Caryophyllaceae) Using Plastid and Homoeologous Nuclear DNA Sequences. *Molecular Phylogenetics and Evolution,* 20**,** 474-481.

RENNY-BYFIELD, S. & WENDEL, J. F. 2014. Doubling down on genomes: polyploidy and crop plants. *American journal of botany,* 101**,** 1711-1725.

ROSATO, M., CASTRO, M. & ROSSELLÓ, J. A. 2008. Relationships of the woody Medicago species (section Dendrotelis) assessed by molecular cytogenetic analyses. *Annals of botany,* 102**,** 15-22.

ROUX, C. & PANNELL, J. R. 2015. Inferring the mode of origin of polyploid species from next-generation sequence data. *Molecular ecology*.

SALMON, A. & AÏNOUCHE, M. 2015. Next-generation sequencing and the challenge of deciphering evolution of recent and highly polyploid genomes. *Next Generation Sequencing in Plant Systematics***,** ?

SANG, T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical reviews in Biochemistry molecular biology and evolution,* 37**,** 121-147.

SANKOFF, D. & ZHENG, C. 2018. Whole Genome Duplication in Plants: Implications for Evolutionary Analysis. *In:* SETUBAL, J. C., STOYE, J. & STADLER, P. F. (eds.) *Comparative Genomics: Methods and Protocols.* New York, NY: Springer New York.

SÉMON, M. & WOLFE, K. H. 2007. Consequences of genome duplication. *Current opinion in genetics development,* 17**,** 505-512.

SMALL, E. 2011. *Alfalfa and relatives*.

SMALL, E. & JOMPHE, M. 1989. A synopsis of the genus Medicago (Leguminosae). Can. J. Bot. 67:3260-3294.

SOLTIS, D., SOLTIS, P. & RIESEBERG, L. H. 1993. Molecular data and the dynamic nature of polyploidy. *Critical reviews in plant sciences,* 12**,** 243-273.

SOLTIS, D. E.*, et al.* 2009. Polyploidy and angiosperm diversification. *American journal of botany,* 96**,** 336-348.

SOLTIS, D. E., BUGGS, R. J., DOYLE, J. J. & SOLTIS, P. S. 2010. What we still don't know about polyploidy. *Taxon,* 59**,** 1387-1403.

SOLTIS, D. E. & SOLTIS, P. S. 1995. The dynamic nature of polyploid genomes. *Proceedings of the National Academy of Sciences of the United States of America,* 92**,** 8089.

SOLTIS, P. S., MARCHANT, D. B., VAN DE PEER, Y. & SOLTIS, D. E. 2015. Polyploidy and genome evolution in plants. *Current opinion in genetics & development,* 35**,** 119-125.

STEBBINS, G. 1947. Types of polyploids: their classification and significance. *Advances in genetics,* 1**,** 403-29.

STEBBINS, G. L. 1969. The significance of hybridization for plant taxonomy and evolution. *Taxon***,** 26-35.

STEBBINS JR, C. 1950. Variation and evolution in plants: Progress During the Past Twenty Years. *In:* HECHT, M. K. & STEERE, W. C. (eds.) *Essays in Evolution and Genetics in Honor of Theodosius Dobzhansky.* Springer, Boston, MA.

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.

WAGNER JR, W. 1970. Biosystematics and evolutionary noise. *J Taxon***,** 146-151.

WANG, K.*, et al.* 2012. The draft genome of a diploid cotton Gossypium raimondii. *Nature genetics,* 44**,** 1098.

WENDEL, J. F. 2015. The wondrous cycles of polyploidy in plants. *American journal of botany,* 102**,** 1753-1756.

WOLFE, K. H. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet*, 2, 333-341.

WU, R., GALLO-MEAGHER, M., LITTELL, R. C. & ZENG, Z.-B. 2001. A General Polyploid Model for Analyzing Gene Segregation in Outcrossing Tetraploid Species. *Genetics,* 159**,** 869-882.

YOUNG, N. D.*, et al.* 2011. The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature,* 480**,** 520-524.

Acknowledgements

In the past four years, I have met many inspiring people that have become good friends and colleagues. I would like to thank everyone for the time I had here at “Botan”, it has been a privilege to do great science together.

I am thankful to:

Supervisor **Alexandre Antonelli** for the opportunity to develop this project and giving me the chance to work with what I am passionate about. There have been bumps in the road with this project, but Alex has provided with great support and advice to get me through.

Former supervisor **Bernard E. Pfeil** for sharing his knowledge on this subject and for the support whenever needed. Even though I found out how interesting this group of plants really are, a little too late.

Bengt Oxelman that stepped in as a supervisor in my last year and provided with great discussions and advice on the projects of my thesis. I cannot thank enough for the support and believing in me when times have been stressful and dire. That you have taken you time to get into the projects and fully understood what I have been working on the past four years. THANKS!

Christine D. Bacon for also stepping in as an adviser/supervisor and for being such an inspiration to all men and women in science. You have dedicated time and enormous support in the last year of my PhD. I thank you for the great moments, climbing, dinners and talks.

Mari Källersjö, examiner, for listening when I needed to talk, for believing in me and my work, and for providing with great advice that have resulted in a finished thesis.

Vivian Aldén for assisting me SO many times in the lab and teaching me all about DNA and RNA extractions. I will share and spread the knowledge you gave to me.

Mum and **Dad**, for the endless and unconditional love you have given me. You have understood the struggles I have encountered and let me "finish talking" when I need to. You gave me the fire that I needed to keep going. So many times, have I failed to describe my project to you, but I hope that this thesis can show you the time and support that you have given me resulted in.

Marcus Eriksson, big brother, for the many hours of talks on the phone during my PhD: the many times of boardgames, movie nights and endless amounts of candy. I only accept Bamse as a gift for my achievements!

Daniel Eriksson, elder brother (not super old, but older) and his family for sheering me on these past years! I am sorry for all the time I have spent in Gothenburg and for the missed birthday parties. I am afraid that it might continue… but when I get rich and successful, I will give it all back!

My cool sister **Anna Koolen**, for being the most vocal member in this family. I have laughed so hard for the emails you have written and you have educated me well in this "WORD" area. I will always be your little tri-cycle sister, reading Pocahontas and tip toing into your room saying "Hej Anna!", but also the crazy scientist you envisioned I become.

Frida Henriksson, best friend! We are both a bit crazy and maybe that is why we go so great together. Think of all the songs we wrote, the theater and weird bug hunting we did. We share the same interests and I do hope we follow our dreams; two farms, a donkey, no money, but happiness in growing vegetables.

Rebecka Åhman, for providing me with a home the first year in Gothenburg. For letting me play with your dogs and cats! For all the walks and the talks in the nice weather. I really thank you!

I thank all the Antonellians that have come and gone! We had great moments and inspiring discussions. None mentioned, none forgotten.

To the colleagues at Botan! You have been the best! You placed the bar high for my next adventure.

To my Kung Fu friends. We had great moments, getting injuries, kicking, jumping and smashing our heads with the bloody sticks! I miss those times.

This project received funding from the Department of Biological and Environmental Sciences and the Faculty of Sciences at the University of Gothenburg, the European Research Council under the European Union's Seventh Framework Programme [FP/2007-2013, ERC Grant Agreement n. 331024], the Knut and Alice Wallenberg Foundation through a Wallenberg Academy Fellowship, the Swedish Research Council (2015-04857), and the Swedish Foundation for Strategic research.