- Title: Phylogenetics of allopolyploids

- Bengt Oxelman*[1] bengt.oxelman@gu.se

- Anne Kragh Brysting[2] a.k.brysting@ibv.uio.no

- Graham R. Jones[1] art@gjones.name

- Thomas Marcussen[3] thmsmrcssn@gmail.com

- Christoph Oberprieler[4] christoph.Oberprieler@biologie.uni-regensburg.de

- Bernard E. Pfeil[1] bernard.pfeil@bioenv.gu.se

- *Corresponding author,

- [1]Gothenburg Global Biodiversity Centre, Department of Biology and Environmental Sciences, University of Gothenburg, Box 461, SE405 30 Göteborg, Sweden

- [2]Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, P.O. Box 1066 Blindern, NO-0316 Oslo, Norway

- [3]Department of Plant Sciences, Norwegian University of Life Sciences, P.O.Box 5003, 1432 Ås, Norway

- [4]Evolutionary and Systematic Botany Group, Institute of Plant Sciences, University of Regensburg, Universitätsstr. 31, D-93053 Regensburg, Germany

Abstract. We give an overview of recently developed methods to reconstruct phylogenies of taxa that includes allopolyploids which have originated in relatively recent times, i.e., such that at least some of the parental lineages of lower ploidy levels are not extinct, and that there is clear ploidy information shown by variation in chromosome counts. We review how these methods have been applied to empirical data. The challenges are discussed and prospects for the future are outlined. The review includes a description of a new version of the AlloppNET method, which now can handle any number of species at the diploid and tetraploid level, and any num,ber of hybridisations. AlloppNET is an extension of the *BEAST model which can handle allotetraploids.

Key words: Multispecies Coalescent Networks, Allopolyploidy, Phylogenetics

INTRODUCTION

Polyploidy, the presence of more than two nuclear genomes in the non-gametic generation of eukaryotes, imposes certain problems to phylogenetic inference. Allopolyploidy, defined here as the merger of the nuclear genomes of two separate lineages into a new lineage where the parental chromosomes continue to form bivalents at meiosis, violates the tree model usually used in phylogenetics, and necessitates a more complicated network approach. There has been a considerable amount of work on hybridisation, often focussed on homoploid hybridisation (e.g., Chen & Wang 2010, 2012, Gerard et al. 2011; Kubatko 2009, Wen et al. 2016). Homoploid hybridisation result in offspring of the same ploidy as the parents. In this review, we focus on the reconstruction of phylogenetic networks where the reticulations are stemming from allopolyploidy.

Approximately 15% of speciation events in angiosperms and 30% in ferns have been estimated to be associated with polyploidy (Wood et al. 2009), and polyploidy is increasingly found also in other eukaryote taxa (Otto & Whitton 2002, Albertin & Marullo 2012). Allopolyploidy has long bheen considered to be the more common mode, but even if allo- and autopolyploidy are considered to be approximately equal in frequency (Barker et al. 2015, see also Doyle & Sherman-Broyles 2017), failure to account for allopolyploidy when reconstructing the past evolution of groups where it has occurred inevitably will lead to inaccurate phylogenetic hypotheses.

In the pre-phylogenetic era, the ancestry of allopolyploids was typically inferred by artificial crossings of potential parental taxa, often coupled with cytological studies of meiotic behaviour, and then comparing the obtained hybrid with the natural polyploid. Potential parental taxa were found by looking for traits from diploids that appeared to having been added up in the polyploid. Although this approach can be viewed as purely inductive, few of these cases have been been tested in a deductive phylogenetic framework, where consideration is also taken to the possibility that the parental lineages may have undergone large changes since the hybridisation event, and even become extinct. It also fails to take into account that some traits may be transgressive, thus being poor predictors of the traits in the parental lineages.

Although the fundamental differences between gene and species trees were acknowledged almost three decades ago (Pamilo & Nei 1988, Doyle 1992), it is only recently that this notion has become an integral part of phylogenetics. In particular, coalescent thoery (Kingman 1982) has been merged with phylogenetics in the multispecies coalescent (MSC) model (Rannala and Yang 2003), which forms a framework where gene trees constitute data for species trees.

In this review, we focus on methodologies used for inferring the evolutionary histories (phylogenies) forming allopolyploids. We pay particular attention to explicit networks, i.e., those where the nodes are thought to represent branching events, or merging events (hybridization), and we review empirical studies where such methodology has been used.

ASSUMPTIONS AND DEFINITIONS

The definitions of auto- and allopolyploidy are somewhat ambiguous in the literature. In the taxonomic definition, autopolyploids arise within species, whereas allopolyploids result from hybridisations between species. Cytogenetic definitions emphasise whether bivalents or multivalents are formed at meiosis or not, with allopolyploids characterised by the former. In reality, auto- and allopolyploids represent the endpoints of a continuum, and many intermediate conditions exist (Soltis et al. 2003). In the methodology presented below, it is generally assumed that recombination between the parental subgenomes has not occurred, or at least not between the studied homoeologues (i.e., genes representing each subgenome). Moreover, in the MSC model, “species” form the branches of the tree, and are assumed to be populations with no selection, random mating, non-overlapping generations, and no migration after divergence, which is abrupt. Thus, these “species” are generally much narrower than most taxonomic species (Toprak et al. 2016, Sukumarana & Knowles 2017).

Although the subgenomes of allopolypolyploids in due course are expected to be subject of recombination, they can remain distinguishable for long periods of time (Renny-Byfield et al. 2014). Therefore, sampling all homoeologous copies of nuclear genes should potentially enable the reconstruction of the genome trees. An obstacle is however that when collecting data from more than one gene, it is usually not possible to a priori assign which gene copies belong to the same subgenome.

DATA ACQUISITION

In this review, we focus on methods aimed at the analysis of molecular sequence data. Several studies have utilised markers such as allozymes (e.g., Roose & Gottlieb 1976, Gastony 1986, Hedrén 1996), microsatellites (e.g., Crespo-López et al. 2007), RAPD/AFLP fragments (e.g., Oxelman 1996, Brochmann et al. 1996, Hedrén et al. 2001), sometimes in concert with gene phylogenies inferred from sequences. However, taken alone, they take a similar approach to the classical morphological studies. That is, additive patterns from putative parental species are used to confirm the ancestry hypothesis. We will focus on methods that use DNA sequences from several species to infer gene and/or species trees or networks.

During a couple of decades, PCR and Sanger sequencing dominated DNA sequencing methodologically. Universal and standardized PCR primers exist for nuclear ribosomal regions, and also for plant plastid and animal/fungal mitochondrial regions, and such regions have dominated phylogenetic research for at least a couple of decades. However, the nuclear ribosomal DNA (nrDNA) cistrons typically undergo concerted evolution in an unpredictable direction, so whereas the plastid tree would be expected to usually reflect the maternal lineage of an allopolyploid, the nrDNA tree could reflect either parent, it could reflect both, or it could be misleading because it is inferred from chimaeric sequences (Álvarez & Mendel 2003).

By using sequences from putatively single-copy nuclear regions, both parental lineages can potentially be recovered. However, although the general phylogenetic utility of such sequences were suggested twenty years ago (reviewed by Sang 2002), the development remained slow until the recent replacement of PCR/Sanger sequencing protocols with Targeted Sequence Enrichment (TSE) and Next Generation Sequencing (NGS) techniques (e.g., Lemmon & Lemmon 2013, Jones & Good 2016). Important reasons for the slow progress are the difficulties in developing universal PCR primers, which often necessitated either knowledge of genomic sequences from the taxonomic group in focus, or modification of primers developed for other taxa to successfully amplify the target regions of the organisms under study. Even if suitable primers are at hand, the problem of disentangling the different homoeologues (with possible allelic variation) into clean sequences remain. Direct sequencing of PCR products is unproblematic for haploid cytoplasmatic loci, and often also employed for nrDNA, in effect resulting in a majority rule consensus sequence from the concerted repeats. If the amplified product contain homoeologues and their possible allelic variants, direct sequencing of PCR products often results in uninterpretable data as a consequence of polymorphic sites and sequence length variation among the amplified copies. The most common approach to circumvent this involves bacterial subcloning of PCR fragments. However, in order to obtain all sequence variants, many sequencing reactions are needed. For example, in order to obtain all sequence variants with 95% probability for two heterozygous homoeologous loci in a tetraploid, at least 14 colonies need to be sequenced (Rautenberg et al. 2008). Alternative, but still rather laborious, strategies involve designing allele-specific primers for secondary sequencing reactions (Scheen et al. 2012), or single molecule PCR (Kraytsberg & Khrapko 2005, Marcussen et al. 2012).

The recent development of TSE/NGS techniques has greatly increased the availability of sequences from many unlinked loci, and holds great promise for the future. In absence of genomic sequence information from the study group, probe design is a problem with similar to PCR primer design, but the scalability of TSE, and also the greedier nature of DNA/DNA-hybridization (Gasc et al. 2016), and the possibility to multiplex barcoded samples with NGS makes it much more efficient. Probe design can be facilitated by using RNA-Seq (e.g., Oxelman et al., unpublished data) or anonymous NGS sequencing such as RADSeq (Davey & Blaxter 2010), genome skimming (Dodsworth 2015), or genotyping by sequencing (Davey et al. 2011). The short read lengths of some sequencing protocols may introduce problems for accurate phasing of alleles/homoeologues/paralogs (Lemmon & Lemmon 2013). Single molecule sequencing techniques hold great promise to mitigate this problem (Rothfels et al. 2017).

AD HOC PHYLOGENETIC METHODS

The first attempts to reconstruct plant allopolyploid origins used gene trees from the plastid genome and nuclear ribosomal regions (often together with other data, e.g.

Oxelman 1996, Brochmann et al. 1996). The approach has been used to infer the phylogenetic origin of the polyploid from its divergent positions in the trees from the two regions, assuming maternal ancestry reflected by the cpDNA tree and paternal from the nrDNA tree, but as noted above this approach leaves many other possibilities.

By using sequences from a putatively single-copy nuclear region, both parental lineages can potentially be recovered, and this approach has been employed by e.g., Small et al. (1998), Sang and Zhang (1999) and Doyle et al. (2000). The gene trees that arise from sampling all homoeologues of a polyploid are multilabeled (MUL-trees), meaning that the same label occurs more than once in the tree.

A problem that arises when sampling homoeologues from several unlinked loci is that it is generally not possible to know a priori which subgenome the sequence belong to. Popp et al. (2005) sampled four RNA polymerase genes, nrDNA ITS, and cpDNA sequences from Arctic di-, tetra- and hexaploid members of *Silene* sect. *Physolychnis*. Using an ad hoc procedure, they presented a “consensus” MUL-tree. Lott et al. (2009a) showed that finding consensus MUL-trees is NP-hard, but presented a heuristic approach implemented in PADRE (Lott et al. 2009b). Huber et al. (2006) devised an exact parsimony method (minimizing the number of reticulation events) to fold MUL-trees into networks. Marcussen et al. (2012) used PADRE to reconstruct an allopolyploid network from a MUL-tree of the GPI gene from *Viola* taxa with ploidy levels spanning from 2x to 18x. For some taxa, the expected number of homoeologues was not recovered. To assess whether this absence was primary, that is, a result of the allopolyploid origin itself, or due to secondary gene loss (or detection failure), the different possible MUL-trees were analysed separately in PADRE and the scenarios were compared to obtain the most parsimonious solution, that is, the one requiring the fewest polyploidisations and gene losses to explain the observed data.

In a second study of *Viola*, Marcussen et al. (2014) identified the most parsimonious network topology from a set of five competing scenarios differing in the interpretation of homoeologue extinctions and lineage sorting, based on the (i) fewest possible ghost subgenome lineages, (ii) fewest possible polyploidization events, and (iii) least possible deviations from the expected ploidy as inferred from available chromosome counts of the involved polyploid taxa. They also estimated the polyploid speciation times by comparing branch lengths and speciation rates of lineages with and without ploidy shifts.

The studies of Marcussen et al. (2012, 2014) show that PADRE has power to deal with high complexity networks. However, a problem with the approach is that it does not take coalescent stochasticity into account. Using sequence data from the angiosperm genus *Fumaria* (ploidy levels ranging between 2x and 14x), Bertrand et al. (2015) proposed a workflow that assigns homoeologues to hypothetical diploid subgenomes prior to genome tree construction. Conflicting assignment hypotheses are evaluated against substitution model error and coalescent stochasticity. Incongruence that cannot be explained by stochastic mechanisms needs to be explained by other processes (e.g., homoploid hybridization or paralogy). The data can then be filtered to build multilabeled genome MUL-trees using inference methods that can recover species trees (e.g., under MSC) in the face of substitution model error and coalescent stochasticity. The network can then be obtained from PADRE. However, another limitation of PADRE is that it is not clear how to incorporate uncertainties in the input MUL-tree.

EXPLICIT SPECIES NETWORK METHODS

Permutation approach using PhyloNet

A relatively simple and fast work-flow pipeline for the reconstruction of species networks in polyploid complexes based on multi-locus gene trees was described by Oberprieler et al. (2016). It uses a permutation strategy and a parsimony-based principle in species-tree reconstruction (Minimising Deep Coalescences, MDC) for the assignment of homoeologs to parental genomes in allopolyploids and eventually constructs a species MUL-tree in which polyploid taxa are represented by their diploid subgenomes. The method utilises the PhyloNet software program (Than et al. 2008, Than & Nakhleh 2009), the first program in which an exact algorithm for inferring species trees from gene trees by minimising the number of extra lineages (Maddison 1997, Maddison & Knowles 2006) had been implemented (Than & Nakhleh 2010).

In a first step, individually for each polyploid accession and each sequenced locus, the procedure forms all possible parental (diploid) allele pairs for the polyploid accession analysed and runs a MDC analysis together with all diploid accessions, which tries to find a species tree that minimises the number of deep coalescences for the given gene tree. For the following permutation steps, *that* combination of parental alleles in the polyploid is kept, for which the lowest number of deep coalescences in the species tree inference procedure is found.

After having determined the optimal allele combinations at all sequenced loci for the polyploid accession under study, a further MDC analysis is run based on all diploid accessions and the single polyploid accession concerned, in which at each turn combinations of allele pairs across loci are submitted to a species tree reconstruction based on all gene trees, repeating this step for all possible combinations of allele groupings across loci. As in the first step of the procedure, *that* allele pair combination across loci is kept that results in a species tree with the minimum number of deep coalescences.

By repeating the two steps described for all polyploid accessions individually, this procedure results in a data set consisting of all diploid accessions (allele pairs) and all pseudo-diploid parental combinations of alleles and allele pairs across loci that could be subjected to a final MDC analysis to reconstruct the overall MUL-species tree. The inference of a species network is then easily accomplished by joining branches representing parental diploids to the polyploids into reticulations using the software programs PADRE (Huber et al. 2006, Lott et al. 2009) or Dendroscope (Huson & Scornavacca 2012).

The permutations necessary for the two steps of the described procedure, along with the handover of re-tailored (pruned) gene tree topologies and further arguments for the MDC search to PhyloNet, is presently implemented in a Matlab v8.0.0.783

(The MathWorks Inc. 2012) script utilising the Matlab Bioinformatics Toolbox (Henson & Cetto 2005). Simulations carried out by Oberprieler et al. (2016) with varying effective population sizes, different temporal scenarios for diploid/polyploid formations, and different types of allopolyploid speciation (sister-species diploids, non-sister diploids, or polyploidisation involving the participation of ancestral diploids) showed that the proposed work-flow pipeline leads to trustworthy reconstructions if effective populations sizes are not large and divergences among ancestral taxa are not shallow.

Simultaneous gene tree and species network inference (AlloppNET)

AlloppNET (Jones et al. 2013) is based on a fully parameterised stochastic model of the relevant evolutionary processes (i.e., the MSC). It uses a Bayesian approach in which parameters are co-estimated by sampling from the posterior distribution using the Markov chain Monte Carlo algorithm. AlloppNet is an extension of the MSC model implemented in *BEAST (Heled & Drummond 2010). Here, we present (see Supplemental Material for details) an extension of the version presented in Jones et al. (2013) that caters for more than one hybridisation and more than two diploid species. It is implemented in BEAST 1.8.3 (Drummond et al. 2012). There is no support in BEAUTi, but R scripts are available (see Supplemental Material) to aid the generation of suitable XML files. AlloppNET can be used with DNA sequence data from diploid species and allotetraploid species. It can handle for multiple individuals per species and multiple unlinked loci. The sequences from the allotetraploid species are assumed to belong to homoeologue pairs for each locus. The identity of the homoeologues (i.e., which sequence came from which diploid species) is estimated along with the other parameters.

It is assumed that all the allotetraploid species in the sample originated via one or more hybridisations each of which may be followed by ordinary speciations of the allotetraploid species. No assumption is made about the fate of the diploid species which hybridized: they may or may not leave descendants in the sample. There is no limit on the number of allotetraploid species or diploid species. The number of hybridizations is estimated.

There is an underlying species network W in the model, which incorporates a topology, node heights, hybridisation times, and population size parameters along each edge. From W, a multi-labeled tree MW can be derived. This is not an arbitrary multi-labeled tree, since for each hybridisation, there is a pair of identical allotetraploid subtrees. Given MW, the homoeologue identities, and the population size parameters, a prior density for the gene trees is calculated using the multspecies coalescent model. Given the gene trees, the likelihood of the sequence data is calculated in the usual way. The posterior is then a product of these terms and priors for all the parameters. The Supplementary Material contains a formal description of the model and further discussion.

In order to sample from the posterior, operators are needed for all the parameters. The operator which changes the number of hybridizations is particularly complicated, since it also changes the number of node heights and the number of population size parameters. This necessitates using a reversible jump move (Green 1995). The SI contains details of the operators, and results for simulated data. Rothfels et al. (2017) used AlloppNET to derive a phylogenetic network in the fern family Cystopteridaceae.

EMPIRICAL EXAMPLES

Galeopsis

The work of Müntzing (1932) represents a milestone in the study of polyploidy as it is the first report of an experimental synthesis of a naturally occurring allopolyploid species, *Galeopsis tetrahit* (Lamiaceae). Müntzing succeeded, via a two-step process involving two diploid *Galeopsis* species and a triploid bridge, to recreate an allopolyploid speciation event producing an allotetraploid plant, which in chromosome number and morphology was very similar to, and cross-compatible with the naturally occurring tetraploid. The *Galeopsis* system provides an example of allopolyploid speciation of a relatively young date, making it relatively easy to use molecular markers to detect both parental contributions and reconstruct the phylogenetic relationships. Even with a single nuclear maker (a non-coding region of NRPA2, encoding the second largest subunit of RNA polymerase I), Bendiksby et al. (2011) confirmed Müntzing's conclusion regarding the allopolyploid origin of the naturally occurring *G. tetrahit* (as well as another naturally occurring tetraploid, *G. bifida*). NRPA2 appeared to be single-copy in *Galeopsis*, and based on an initial nested PCR procedure, degenerate primers (Popp and Oxelman 2004) and subcloning of the products, parental-homoeolog specific primers were developed for direct sequencing. Two diverged NRPA2 copies were found in each of the two tetraploids, with the resulting MUL-tree clearly demonstrating that both tetraploids originated by allopolyploid speciation from the diploid *G. speciosa* and *G. pubescens* lineages. However, the data show that the parental genomes involved in the two tetraploids differed genetically and that the two most likely originated by independent polyploidisation events. The addition of cpDNA markers further made it possible to determine the maternal parent of *G. tetrahit*. Bendiksby et al. (2011) also analysed a larger population sampling of the two tetraploids and their parental species using amplified fragment length polymorphisms (AFLPs). These results, in combination with the DNA sequence data, suggest that both tetraploids appear to have originated only once, as opposed to recurrent origins usually reported for natural polyploids (e.g., Soltis et al. 2003), and that frequent hybridization and introgression takes place especially within ploidy levels.

Cerastium

As a typical example of recent and rapid speciation during the Pleistocene, the high-ploid *Cerastium alpinum* group (Caryophyllaceae) represents a much more complex polyploid history. The complex consists of six high-ploid species (octoploids, 2n = 8x = 72, and dodecaploids, 2n = 12x = 108) with their main distribution in arctic or alpine regions, and for which no diploid progenitors are known. Brysting et al. (2007, 2011) sequenced non-coding regions of three single-copy nuclear RNA polymerase genes. The sequences were merged into consensus sequences representing monophyletic groups in initial phylogenetic analyses and used to produce multilabeled trees. The multilabeled gene trees were transformed into networks using the PADRE software (Lott et al. 2009b). The closest living relatives of the *C. alpinum* group are tetraploid species (2n = 36). Despite this, only one functional copy of each of the three genes was detected and these tetraploid taxa are most likely the result of ancient polyploidisation events. The high-ploid species of the *C. alpinum* group are on the other hand likely results from much more recent polyploidisation events related to recurrent episodes of range expansions and contractions during the Quaternary glaciations, and in most cases the copy number corresponds well with ploidy level. Overall, Brysting et al. (2007, 2011) were successful in disentangling the tetraploid progenitor lineages of the high-ploid species of the *C. alpinum* group. However, the three networks based on different Pol regions differed in several aspects and had small deviations from the general pattern which could better be explained by gene duplication, lineage sorting events or lack of information by incomplete sampling. The fact that gene loss, pseudogenisation and possible lineage sorting are working independently in different parts of the polyploid genome may hamper the interpretation of reticulate evolution even in relative young plant groups such as the *C. alpinum* complex, and emphasises the importance of approaches where several unlinked regions of the genome are examined.

Arabidopsis

The two tetraploid species *Arabidopsis suecica* and *A. kamchatica* both have allopolyploid origins involving diploid *A. thaliana* and *A. arenosa* progenitors in the first case (Comai et al. 2000, Jakobsson et al. 2006) and diploid *A. lyrata* and *A.*

halleri ssp. *gemmifera* in the second case (Shimizu-Inatsugi et al. 2009, Jørgensen 2011). It is assumed that *A. suecica* originated from a single Pleistocene speciation event dating back to between 20,000 and 300,000 years ago with a single origin and a present day limited distribution in Fennoscandia (Jakobsen et al. 2006). For *A. kamchatica*, the combination of nuclear ITS and plastid trnL-F sequences (Schmickl et al 2010) and low-copy nuclear markers (Shimizu-Inatsugi et al. 2009, Jørgensen 2011) suggest that this allopolyploidisation has happened several times (and in several places), or alternatively that there has been subsequent gene flow from the parents to the descendant.

Arabidopsis arenosa and *A. lyrata* both exist in diploid and tetraploid forms. At the diploid level these two species seem to be reproductively isolated since they are genetically and phenotypically distinct and no past or recent gene flow can be detected between them from low-copy nuclear sequences (Jørgensen 2011) nor by a combination of microsatellite and cpDNA markers (Schmickl & Koch 2011). Reproductive incompatibility at the diploid level has also been confirmed by reciprocal crosses (Muir et al. 2015). Low-copy nuclear sequences in combination with cpDNA sequences, however, showed that there is interspecific gene flow between tetraploids of the two species, and also bidirectional gene flow across ploidy levels within each species (Jørgensen 2011). Polyploidization thus may play a prominent role in plant speciation, not only by mere reproductive isolation but also by loosening hybridization barriers between related species.

Recently Novikova et al. (2016) resequenced the genomes of 94 individuals covering all *Arabidopsis* taxa (including subspecies) and sampled throughout the entire geographic range. In total, 9,119 genes were used for genus-wide alignments for 25 Mb of the genome. Clustering (Neighbour Joining and Admixture) on the basis of genome-wide polymorphisms was in agreement with previous morphological and molecular results revealing four major groups, which corresponded to the widely distributed species *A. thaliana, A. halleri, A. lyrata* and *A. arenosa*, as well as three minor groups, corresponding to the geographically limited *A. croatica, A. cebennensis* and *A. pedemontana*. The few included samples of the allopolyploids *A. suecica* and *A. kamchatica* came out as expected, with sequences grouping together with both of the parental species. At the level of individual gene trees (from Maximum likelihood analyses), 100% supported the monophyly of *A. thaliana*, which has been estimated to have diverged from the other taxa at least 6 mill. years ago (Hohmann et al. 2015). None of the remaining taxa were universally supported by all gene trees and overall shared polymorphisms demonstrated that reproductive isolation has been considerably more recent than previously estimated divergence times. The data suggest multiple cases of past gene flow between all species that contradict a bifurcating species tree despite the fact that several of the current species of the *Arabidopsis* genus are clearly identifiable morphologically or by polymorphism data, and Novikova et al. (2016) question the utility of the tree as model for speciation.

Medicago

Despite recent developments, some phylogenetic problems are still extremely difficult to solve because of complications arising from specific properties of the biological system under study. Polyploids arising from within (or among close relatives of) the *Medicago sativa* (Fabaceae) complex are among such cases because of the complexity of historical interactions among this group of species.

Small (2011) recognises eight subordinate taxa within *Medicago sativa*, including the world's most important forage crop, alfalfa (*M. sativa ssp. sativa*). Hybridisation has resulted in the successful introgression of alleles between subspecies of the complex at the same ploidy level (Sakiroglu et al. 2010, Kaljund & Leht 2013). Tetraploids within the complex are presumed to have arisen via autotetraploidy (Havananda et al. 2011), with alfalfa known to display tetrasomic inheritence (McCoy & Bingham 1988). Unreduced gamete production in diploids, and crosses between 2x and 4x plants to produce viable 4n plants, has been demonstrated experimentally (Bingham 1968; Veronesi et al. 1986) and probably allows gene flow from 2x to 4x populations to occur in the wild. Together, these lines of evidence have been used to propose a model of the *M. sativa* complex comprising four pillars (Small 2011), each pillar being a diploid-autotetraploid taxon pair with similar morphologies. Gene flow is readily achieved among diploid and among tetraploid members of each pillar and also occurring to some degree from diploid to tetraploid members (Small 2011).

<u>Medicago arborea</u> and <u>M. strasseri</u> are both tetraploids and morphologically very similar to one another and to <u>M. citrina</u>, a hexaploid; all three of them are placed in the same section (Small 2011). *Medicago citrina* is thought to be an autohexaploid species (Quiros & Bauchan 1988), presumably based on the morphological similarity, although this has been questioned on the basis of a lack of clear chromosomal similarities (Rosato et al., 2008). If the former case is correct, then this implies that *M. arborea* and *M. strasseri* are also autotetraploids, a hypothesis explicitly made in an earlier study, on the basis that no known diploid forms resembling *M. arborea* were known (Lesins & Lesins 1979). *Medicago arborea*, as a representative of these three closely related woody species, was also considered to be the "oldest" member of the genus, and therefore not a likely candidate to be an "immediate progenitor" of other perennial species related to the *M. sativa* complex (Lesins & Lesins

1979, Quiros & Bauchan 1988). This reasoning was based at least in part on the presumption that woodiness is an ancestral trait in *Medicago* (Lesins & Lesins 1979), as it is in other groups of plants. However, both the assumption of *M. arboera*'s phylogenetic position and the kind of polyploid origin have been called into question by a recent molecular systematic study using 10 nuclear genes (Eriksson et al. 2017). In this work, phylogenetic evidence points strongly towards an allopolyploid origin, because clades of alleles that presumably represent homoeologues are not sister in eight out of 10 gene trees (Eriksson et al. in prep.). The conclusion is also supported by AlloppNET, but homoploid hybridisations at the diploid level are likely violating the assumptions of the model, and results in slow convergence of the MCMC. In this case at least, understanding of the polyploid origin is progressing, even though the relationships among diploids is contradictory across gene trees (e.g., de Sousa et al. 2016, Eriksson et al. 2017).

FUTURE CHALLENGES AND PROSPECTS

Our understanding of polyploidy is rapidly increasing, and the processes as well as the patterns produced are intricate (Barker et al. 2016, and references therein). It is clear that polyploidy has played, and is playing, a central role in the evolution of some groups, obviously plants, but also in fungi, invertebrates, vertebrates, and some other eukaryotes. It is therefore of uttermost importance to have tools to reconstruct the phylogenetic past. Phylogenetics has developed into a high level of sophistication, and models that can take coalescent stochasticity into account are becoming standard (Edwards & Rausher 2009, Degnan & Rosenberg 2009). Recent advances have also made it possible to use these models without prior knowledge about species delimitations (e.g., Jones et al. 2014, Toprak et al. 2016). However, it is necessary to take homoploid and polyploid hybridisation into account. We also need to develop better procedures to identify paralogs, a problem that may be overlooked, and especially prevalent in groups where polyploidy, ancient and modern, is common. And in the end, models that are able to take all relevant parameters into account are desirable.

We have reviewed some of the recent progress of methods to infer phylogenetic networks in presence of allopolyploids, and although the advances are promising, much more efforts are needed in order to meet the information inherent in the enormous data sets now being generated with NGS. As in phylogenetics in general, we see stepwise approaches that first estimate gene trees, which are then used as data for species tree inference, or simultaneous co-estimation of gene and species trees and their parameters. There is also a dichotomy in methods that first seek to infer a genome MUL-tree, which then may be converted to a network, and methods that infer the network directly. Fully parameterised, direct methods are so far restricted to two ploidy levels, but are perhaps the most promising. On the other hand, they are sensitive to some model violations, which may comprimise convergence of MCMC, so simpler methods have their justification as well.

LITERATURE CITED

Albertin W, Marullo P. 2012. Polyploidy in fungi: Evolution after whole-genome duplication. *Proc. R. Soc. B. Biol. Sci.* 279:2497–509

Álvarez I, Wendel JF. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Mol. Phylogen. Evol.* 29:417–434

Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA. 2015. On the relative abundance of autopolyploids and allopolyploids. *New Phytol.* 210:391–398

Barker MS, Husband BC, Pires JC. 2016. Spreading Winge and flying high: The evolutionary importance of polyploidy after a century of study. *Am. J. Bot.* 103:11391145

Bendiksby M, Tribsch A, Borgen L, Trávníček P, Brysting A.K. 2011. Allopolyploid origins of the *Galeopsis* tetraploids - revisiting Müntzing's classical textbook example using molecular tools. *New Phytol.* 191:1150-1167

Bertrand YJK, Scheen A-C, Marcussen T, Pfeil BE, de Sousa F, Oxelman B. 2015. Assignment of Homoeologues to Parental Genomes in Allopolyploids for Species Tree Inference, with an Example from *Fumaria* (Papaveraceae) *Syst. Biol.* 64:448-471

Bingham ET. 1968. Transfer of diploid *Medicago* spp. germplasm to tetraploid *M. sativa* L. in 4x-2x crosses. *Crop Sci.* 8:760-762.

Brochmann C, Nilsson T, Gabrielsen TM. 1996. A classic example of postglacial allopolyploid speciation re-examined using RAPD markers and nucleotide sequences: *Saxifraga osloensis* (Saxifragaceae). *Symb. Bot. Ups.* 31:75–89

Brysting AK, Oxelman B, Huber KT, Moulton V, Brochmann C. 2007. Untangling complex histories of genome merging in high polyploids. *Syst. Biol.* 56:467–476.

Brysting AK, Mathiesen C, Marcussen T. 2011. Challenges in polyploid phylogenetic reconstruction: A case story from the arctic-alpine *Cerastium alpinum* complex. *Taxon* 60: 333-347

Chen Z-Z, Wang L. 2010. Hybridnet: a tool for constructing hybridization networks. *Bioinfor- matics* 26:2912–2913

Chen Z-Z, Wang L. 2012. Algorithms for reticulate networks of multiple phylogenetic trees. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 9:372–384

Comai L, Tyagi AP, Winter K, Holmes-Davis R, Reynolds SH, Stevens Y, Byers B. 2000. Phenotypic instability and rapid gene silencing in newly formed Arabidopsis allotetraploids. *Plant Cell* 12:1551-1568.

Crespo-López ME, Pala I, Duarte TL, Dowling TE, Coelho MM. 2007. Genetic structure of the diploid-polyploid fish *Squalius alburnoides* in southern Iberian basins Tejo and Guadiana, based on microsatellites. *J. Fish Biol.* 71:423–436

Davey JW, Blaxter ML. 2010. RADSeq: next-generation population genetics. *Brief. Funct. Genomics* 9:416–423

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen, JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Rev. Gen.*12:499-510

de Sousa F, Bertrand YJK, Pfeil BE. 2016. Patterns of phylogenetic incongruence in *Medicago* L. found among six linkage groups. Pl. Syst. Evol. 302:493–513

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.*. 24:332-340

Dodsworth S. 2015. Genome skimming for next-generation biodiversity analysis. *Trends Pl. Sci.* 20:525–527

Doyle JJ. 1992. Gene trees and species trees - molecular systematics as one-character taxonomy. *Syst. Bot.* 17:144-163

Doyle JJ, Doyle JL, Brown AHD, Pfeil BE. 2000. Confirmation of shared and divergent genomes in the *Glycine tabacina* polyploid complex (Leguminosae) using histone H3–D sequences. *Syst. Bot*. 25:437–448

Doyle JJ, Sherman-Broyles S. 2017. Double trouble: taxonomy and definitions of polyploidy.

New Phytol. 213:487–493

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29: 969-1973

Edwards SV, Rausher M. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63:1-19

Eriksson JS, Blanco Pastor JL, Sousa F, Bertrand YJK, Pfeil BE. 2017. A cryptic species produced by autopolyploidy and subsequent introgression involving *Medicago prostrata* (Fabaceae). *Mol. Phyl. Evol.* 107:367-381

Gasc C, Peyretaillade E, Peyret P. 2016. Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Res.* 44:4504-4518

Gastony GJ. 1986. Electrophoretic evidence for the origin of fern species by unreduced spores. *Amer. J. Bot.* 73:1563-1569

Gerard D, Gibbs HL, Kubatko LS. 2011. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. *BMC Evol. Biol.* 11:291

Green PJ. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82:711–732

Havananda T, Brummer EC, Doyle JJ. 2011. Complex patterns of autopolyploid evolution in alfalfa and allies (*Medicago sativa*; Leguminosae). *Am. J. Bot.* 98:1633-1646.

Hedrén M. 1996. Genetic differentiation, polyploidization and hybridization in northern European *Dactylorhiza* (Orchidaceae): evidence from allozyme markers. *Pl. Syst. Evol.* 201:31-55

Hedrén M, Fay MF, Chase MW. 2001. Amplified fragment length polymorphisms (AFLP) reveal details of polyploid evolution in *Dactylorhiza* (Orchidaceae). *Amer. J. Bot.* 88:1868-1880

Heled J, Drummond AJ 2010. Bayesian Inference of Species Trees from Multilocus Data. *Mol. Biol. Evol.* 27: 570–580

Henson R, Cetto L. 2005. The MATLB bioinformatics toolbox. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics* 4:4.8:105.

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

Huson DH, Scornavacca C. 2012 Dendroscope 3 – An interactive viewer for rooted phylogenetic trees and networks. *Syst. Biol.* 61:1061-1067

Jakobsson M, Hagenblad J, Tavaré S, Säll T, Halldén C, Lind-Halldén C, Nordborg M. 2006. A unique recent origin of the allotetraploid species *Arabidopsis suecica*: Evidence from nuclear DNA markers. *Mol. Biol. Evol.* 23: 1217-1231.

Jones G, Sagitov S, Oxelman B. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Syst. Biol.* 62:467-478

Jones G, Aydin Z, Oxelman B. 2014. DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics* 31:991-998

Jones MR, Good JM. 2016. Targeted capture in evolutionary and ecological genomics. *Mol. Ecol* 25:185–202

Jørgensen MH. 2011. A poly approach to ploidy: polyploid evolution and taxonomic implications. PhD thesis, University of Oslo (http://urn.nb.no/URN:NBN:no-31169)

Kaljund K, Leht M. 2013. Extensive introgressive hybridization between cultivated lucerne and the native sickle medic (*Medicago sativa* ssp. *falcata*) in Estonia. *Ann. Bot. Fenn.* 50:23-31.

Kingman JFC. 1982. On the genealogy of large populations. *J. Appl. Probab.* 19A: 27-43

Kraytsberg Y, Khrapko K. 2005. Single-molecule PCR: an artifact-free PCR approach for the analysis of somatic mutations. *Expert Rev. Mol. Diagn.* 5:809–815

Kubatko LS. 2009. Identifying hybridization events in the presence of coalescence via model selection. *Syst. Biol*. 58:478–488

Lemmon EM, Lemmon AR. 2013. High-throughput genomic data in systematics and phylogenetics. *Ann. Rev. Ecol. Evol. Syst.* 44:1–23

Lesins KA, Lesins I. 1979. Genus *Medicago* (Leguminosae). A taxogenetic study. The Hague: Dr. W. Junk

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

Lott M, Spillner A, Huber KT, Moulton V. 2009b. PADRE: a package for analyzing and displaying reticulate evolution. *Bioinformatics* 25:1199–1200

Maddison WP. 1997. Gene trees in species trees. *Syst. Biol.* 46: 523-536.

Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55: 21-30.

Marcussen T, Jakobsen K, Danihelka J, Ballard H, Blaxland K, Brysting A, Oxelman B. 2012. Inferring Species Networks from Gene Trees in High-polyploid North American and Hawaiian Violets (*Viola*, Violaceae). *Syst. Biol.* 61:107-126

Marcussen T, Heier L, Brysting AK, Oxelman B, Jakobsen K. 2014. From Gene Trees to a Dated Allopolyploid Network: Insights from the Angiosperm Genus *Viola* (Violaceae) *Syst. Biol.* 64:84-101

McCoy TJ, Bingham ET. 1988. Cytology and cytogenetics of alfalfa. *Agron Monograph* 29:737-776

Muir G, Ruiz-Duarte P, Hohmann N et al. 2015. Exogenous selection rather than cytonuclear incompatibilities shapes asymmetrical fitness of reciprocal *Arabidopsis* hybrids. *Ecol. Evol.* 5:1734–1745.

Müntzing A. 1932. Cytogenetic investigations on synthetic *Galeopsis tetrahit. Hereditas* 16:105-154

Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, et al. 2016. Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics* 48:1077-1082

Oberprieler C, Wagner F, Tomasello S, Konowalik K. 2016. A permutation approach for inferring species networks from gene trees in polyploid complexes by minimising deep coalescences. – *Methods Ecol. Evol.* DOI: 10.1111/2041-210X.12694

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu. Rev. Genet.* 34:401–437

Oxelman B. 1996. RAPD patterns, nrDNA ITS sequences, and morphological patterns in the *Silene sedoides*-group (Caryophyllaceae). *Pl. Syst. Evol.* 201:93-116

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5:568-583

Popp M, Oxelman B. 2004. Evolution of a RNA polymerase gene family in *Silene* (Caryophyllaceae) – incomplete concerted evolution and topological congruence among paralogues. *Syst. Biol.* 53: 914–932

Popp M, Erixon P, Eggens F, Oxelman B. 2005. Origin and evolution of a circumpolar polyploid species complex in *Silene* (Caryophyllaceae). *Syst. Bot.* 30: 302-313

Quiros CF, Bauchan GR. 1988. The genus *Medicago* and the origin of the *Medicago sativa* complex. *Agron. Monograph* 29: 737-776.

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

Rautenberg A, Filatov D, Svennblad B, Heidari N, Oxelman, B. 2008. Conflicting phylogenetic signals in the SlX1/Y1 gene in *Silene*. *BMC Evol. Biol.* 8:299

Renny-Byfield S, Gallagher JP, Grover CE, Szadkowski E, Page JT, Udall JA, Wang X, Paterson AH, Wendel JF. 2014. Ancient gene duplicates in *Gossypium* (cotton) exhibit near-complete expression divergence. *Genome Biol. Evol.* 6:559–571

Roose ML, Gottlieb LD. 1976. Genetic and biochemical consequences of polyploidy in *Tragopogon. Evolution* 30:818-830

Rosato M, Castro M, Rosselló JA. 2008. Relationships of the woody *Medicago* species (section *Dendrotelis*) assessed by molecular cytogenetic analyses. *Ann. Bot.* 102:15-22

Rothfels CJ, Pryer KM, Li FW. 2017. Next generation polyploid phylogenetics: rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing. *New Phytol.* 213:413–429

Sakiroglu M, Doyle JJ, Brummer EC. 2010. Inferring population structure and genetic diversity of broad range of wild diploid alfalfa (*Medicago sativa* L.) accessions using SSR markers. *Theor. Appl. Genet.* 121:403-415

Sang T, Zhang D. 1999. Reconstructing hybrid speciation using sequences of low-copy nuclear genes: hybrid origins of five *Paeonia* species based on Adh gene phylogenies. *Syst. Bot*. 24: 148–163

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

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

Schmickl R, Jørgensen MH, Brysting AK, Koch MA. 2010. The evolutionary history of the *Arabidopsis lyrata* complex: A hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evol. Biol.* 10: 98

Schmickl R, Koch MA. 2011. *Arabidopsis* hybrid speciation processes. *Proc. Natl. Acad. Sci. USA* 108:14192–14197

Shimizu-Inatsugi R, Lihová J, Iwanaga H, Kudoh H, Marhold K, Savolainen O, Watanabe K, Yakubov VV, Shimizu KK. 2009. The allopolyploid *Arabidopsis* *kamchatica* originated from multiple individuals of *Arabidopsis lyrata* and *Arabidopsis halleri*. *Mol. Ecol.* 18:4024-4048.

Small E. 2011. Alfalfa and relatives: evolution and classification of *Medicago*. Ottawa: NRC Research Press

Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF. 1998. The tortoise and the hare: choosing between noncoding plastome and nuclear Adh sequences for phylogeny reconstruction in a recently diverged plant group. *Am. J. Bot.* 85:1301–1315

Soltis DE, Soltis PS, Tate JA. 2003. Advances in the study of polyploidy since Plant Speciation. *New Phytol.* 161:173–191

Sukumarana J, Knowles LL. 2017. Multispecies coalescent delimits structure, not species. *Proc. Natl. Acad .Sci. USA* http://www.pnas.org/content/early/2017/01/25/1607921114

Than C, Nakhleh L. 2009. Species tree inference by minimizing deep coalescences. *PLoS Comp. Biol. 5*: e1000501.

Than C, Nakhleh L. 2010. Inference of parsimonious species tree from multilocus data by minimizing deep coalescences. In: Knowles LL, Kubatko LS (eds.), *Estimating Species Trees: Practical and Theoretical Aspects*, Pp. 79-97. – John Wiley & Sons, Hoboken (NJ).

Than C, Ruths D, Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9: 322.

Toprak Z, Pfeil BE, Jones G, Marcussen,T, Ertekin AS, Oxelman B. 2016. Species delimitation without prior knowledge: DISSECT reveals extensive cryptic speciation in the *Silene aegyptiaca* complex (Caryophyllaceae). *Molec. Phyl. Evol.* 102:1-8

Veronesi F, Mariani A, Bingham ET. 1986. Unreduced gametes in diploid *Medicago* and their importance in alfalfa breeding. *Theor. Appl. Genet.* 72:37-41

Wen D, Yu Y, Nakhleh, L. 2016. Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent *PLoS Genetics*, 12:e1006006

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad .Sci. USA* 106:13875–13879